Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Escola Politècnica Superior

Using external sources of bilingual information for word-level quality estimation in translation technologies

Miquel Esplà Gomis

Tesi presentada per a aspirar al grau de
DOCTOR O DOCTORA PER LA UNIVERSITAT D'ALACANT
MENCIÓ DE DOCTOR O DOCTORA INTERNACIONAL
DOCTORAT EN APLICACIONS DE LA INFORMÀTICA

Dirigida per
Dr. Felipe Sánchez Martínez
Dr. Mikel L. Forcada Zubizarreta

Aquesta tesi ha estat finançada pel Ministeri d'Economia i Competitivitat a través dels projectes TIN2009-14009-C02-01 i TIN2012-32615, la Unió Europea a través del projecte Abu-MaTran en virtut de l'acord amb número de concesió PIAP-GA-2012-324414 del Seté Programa Marc FP7/2007-2013 i la Universitat d'Alacant mitjançant el projecte KAZAKHUNIVERSITY1-15I.

*"El pensament s'afirma —i s'aferma— en les objeccions.
Doneu-me un bon contradictor, i seré capaç d'inventar les més excelses teories."*
Joan Fuster

Agraïments

Primerament, vull donar les gràcies als meus directors de tesi, Mikel L. Forcada i Felipe Sánchez Martínez, per haver-me guiat en el procés de recerca que culmina en aquesta tesi doctoral. Mikel i Felipe han sabut supervisar-me, aconsellar-me i motivar-me en aquest procés amb esforç i paciència, cosa que els agraïsc infinitament, però també s'han esforçat per transmetre'm la seua ètica i el seu amor per la recerca i per la faena ben feta, cosa que per a mi no té preu. Gràcies als dos.

Així mateix, vull donar les gràcies a Juan Antonio Pérez Ortiz i a Rafael Carrasco per donar-me l'oportunitat de treballar i aprendre amb ells durant els primers anys dels meus estudis de doctorat a la Universitat d'Alacant. En aquest sentit, també he d'agrair als companys de Prompsit Language Engineering i a les persones que han participant al projecte *Abu-MaTran*, dels quals he pogut aprendre molt, especialment pel que fa a l'adquisició de corpus paral·lels; a més de compartir els seus coneixements amb mi, els dec un agraïment addicional a Gema, Sergio, Antonio, Nikola, Prokopis, Vassilis, Rapha, Tommi i Andy per permetre que dos dels articles que he escrit amb ells formen part d'aquesta tesi.

Tinc un record molt bo dels mesos que vaig passar a Trento amb els companys del grup *HLT* de la *Fondazione Bruno Kessler*. He d'agrair especialment a Matteo Negri la seua supervisió i consell durant aquest temps, així com a José C. de Souza, amb qui va ser un plaer treballar i em va fer sentir com a casa.

També vull agrair als amics que han passat aquests anys pel laboratori i que han fet molt més agradables les llargues jornades de faena, i en especial a Víctor, Xavi, Fran i Daniel, amb els quals he pogut, a més, treballar i compartir idees que m'han fet avançar en la recerca. Gràcies, també, a Majo, per tirar-me una mà amb els dubtes lingüístics d'última hora. De la mateixa forma, vull donar les gràcies a tots els bons amics i amigues (no gose dir tots els noms perquè estic segur de que me'n deixaria a algú que es mereix ser-hi) que m'han animat durant aquests anys i amb qui he aprés coses tan importants —si no més— que les que he aprés a la Universitat. Tota la meua estima i gratitud per a ells.

Aquesta tesi no haguera estat possible sense el suport i l'estima de la meua família durant tots aquests anys. No me n'oblide dels que ja no hi són: un trosset d'aquest treball és també seu. He d'agrair especialment al meu germà Pepe els seus ànims, i als meus pares, Marién i Paco, pel seu ajut i exemple inestimables, per ensenyar-me a valorar la importància del coneixement i per inculcar-me els valors de l'esforç i la perseverança. A ells els ho dec tot.

Bona part de la culpa que aquesta tesi doctoral veja la llum és de Luisa, que m'ha ajudat moltíssim, sobretot durant els darrers mesos d'escriptura. Gràcies per celebrar amb mi els bons moments i per animar-me en els dolents; gràcies per donar-los sentit a tots plegats.

Suport institucional

Vull donar les gràcies a les institucions que han finançat la meua investigació: la Universitat d'Alacant mitjançant el projecte KAZAKHUNIVERSITY1-15I; el Ministeri d'Economia i Competitivitat a través dels projectes TIN2009-14009-C02-01 i TIN2012-32615; i la Unió Europea a través del projecte Abu-MaTran (núm. de concesió PIAP-GA-2012-324414) del Seté Programa Marc FP7/2007-2013. Així mateix, vull agrair a les empreses Reverso-Softissimo i Prompsit Language Engineering per permetre'm l'accés a l'eina Reverso Context i al *University Research Program for Google Translate* per permetre'm utilitzar el seu servei *Google Translate* per als experiments descrits en aquesta tesi.

Miquel Esplà Gomis
Alicante, 9 de desembre de 2015

Resum en català

L'estimació de la qualitat de la traducció (EQ) consisteix a predir el nivell de qualitat d'una traducció en llengua meta (LM) produïda per a un segment en llengua origen (LO). L'EQ ha esdevingut crucial per a les tecnologies de la traducció: els traductors necessiten una EQ acurada per a predir l'esforç requerit en una tasca de traducció i per a escollir la tecnologia de traducció a utilitzar. Aquesta tesi doctoral descriu una col·lecció de noves tècniques per a l'EQ de dues tecnologies de traducció: la traducció automàtica (TA) i la traducció assistida per ordinador (TAO) basada en memòries de traducció (MT).

Els mètodes proposats usen qualsevol font d'informació bilingüe (FIB) disponible de manera agnòstica, és a dir, sense fer cap mena d'assumpció pel que fa a la quantitat, la qualitat, o el format de la informació bilingüe utilitzada. En el context d'aquesta tesi doctoral, s'anomena FIB a qualsevol recurs capaç de proporcionar traduccions en una llengua per a un subsegment[1] donat en una altra llengua. Per poder aplicar els mètodes desenvolupats a parells de llengües amb pocs recursos bilingües, part de la recerca s'ha dedicat a l'adquisició de FIB d'Internet.

L'objectiu d'aquesta introducció és presentar els conceptes bàsics sobre l'EQ per a les tecnologies de la traducció, presentar la motivació de la recerca desenvolupada i posar els diferents articles reimpresos inclosos en aquesta tesi en un marc comú.

Objectius i resultats de la tesi

L'objectiu principal d'aquesta tesi doctoral és desenvolupar mètodes per a l'EQ, fent servir FIB, tant per a TA, com per a TAO basada en MT. La motivació de les tecnologies que es descriuen en aquesta tesi és aprofitar les FIB existents que són disponibles, per exemple, a Internet, com ara els diccionaris bilingües, les taules de

[1] Un subsegment és una seqüència d'una o més paraules contigües que formen part d'una oració.

subsegments o "frases",[2] la TA, les MT, o els cercadors de concordances bilingües. La *hipòtesi de treball principal* d'aquesta tesi és la següent:

Hipòtesi de treball principal: És possible desenvolupar mètodes exclusivament basats en FIB externes per a estimar la qualitat de la traducció de cada mot, tant en TA com en TAO basada en MT.

Aquesta hipòtesi de treball sintetitza els objectius principals del treball desenvolupat en aquesta tesi doctoral, i proporciona un fil conductor per a descriure'l. La recerca duta a terme per a confirmar aquesta hipòtesi de treball es divideix en tres blocs:

- desenvolupament de mètodes basats en FIB per a l'EQ de cada mot en TAO basada en MT;
- desenvolupament de mètodes basats en FIB per a l'EQ de cada mot en TA; i
- desenvolupament de mètodes per a l'obtenció de FIB per a parells de llengües amb pocs recursos.

Aquesta secció té per objectiu descriure els problemes que s'han abordat al llarg d'aquesta tesi doctoral i les solucions proposades per a cadascun d'ells.

Cal emfatitzar que les tècniques per a l'EQ de cada mot desenvolupades en aquesta tesi són agnòstiques pel que fa a les FIB utilitzades; això garanteix que els mètodes resultants siguen flexibles i que, per tant, s'aprofiten al màxim les FIB.

Ús de fonts d'informació bilingües per a l'estimació de la qualitat de la traducció en traducció assistida per ordinador basada en memòries de traducció

L'objectiu d'aquest bloc de recerca és definir mètodes per a l'EQ de cada mot per a TAO basada en MT. Les eines de TAO basada en MT funcionen de la següent manera: quan el traductor vol traduir un nou segment S' en LO, l'eina cerca a la MT les unitats de traducció (S, T) amb un segment en LO S semblant a S' i les presenta a l'usuari com a suggeriments de traducció. D'aquesta forma, el segment corresponent en LM T pot ser utilitzat com a punt de partida per a traduir S'.

[2] Les taules de sub-segments, en anglés *phrase tables*, són un component intern dels sistemes de TA estadística basada en subsegments (Koehn et al., 2003). Bàsicament, són memòries de traducció que contenen parells de subsegments que són traduccions mútues i que s'extrauen mitjançant l'alineament de les paraules de corpus paral·lels (Och and Ney, 2003).

Per saber com són de semblants un segment S i S', les eines de TAO basades en MT usen *mètriques de concordança parcial* — anomenades *fuzzy-match score* en la bibliografia en anglés (Sikes, 2007). Tot i que existeix una àmplia varietat de mètriques de concordança parcial, la gran majoria es basen en algorismes de distància d'edició (Levenshtein, 1966) en què es comparen els mots de dues cadenes. Aquestes mètriques solen presentar-se al traductor en forma de percentatges per facilitar l'estimació de l'esforç requerit en posteditar un suggeriment de traducció. Així, una concordança parcial del 100% indica que els segments S i S' són idèntics i que, per tant, el segment T en LM podria ser utilitzat com a traducció de S' sense fer-hi cap edició. Per contra, una concordança parcial del 0% implicaria que S i S' no s'assemblen gens i que, per tant, el segment T en LM no ajudaria gens en la traducció de S'. Les mètriques de concordança parcial esdevenen, per tant, mètriques d'estimació de la qualitat dels segments traduïts. De fet, és habitual que els mots de S que no concorden amb S' siguen destacades a l'hora de presentar al traductor els suggeriments de traducció; tanmateix aquesta informació no es proporciona per a la LM, on esdevindria molt més útil. L'objectiu del treball desenvolupat en aquest bloc de recerca és anar un pas més enllà i projectar la informació dels mots en S que no concorden amb S' sobre T, per a obtenir una EQ de cada mot.

És obvi que proporcionar una EQ sobre T seria molt més informatiu a l'hora d'estimar l'esforç requerit per completar la tasca de traducció. A més, si aquesta informació fóra presentada al traductor seria possible guiar-lo en la tasca de postedició. Per exemple, els mots que han de ser modificats (eliminats o substituïts) podrien ser acolorits en roig, mentre que els mots que poden romandre tal com estan, podrien ser acolorits en verd.

Malgrat els avantatges de l'EQ de cada mot, l'única referència a aquesta tasca en la bibliografia és la patent de Kuhn et al. (2011). Lògicament, pel fet de tractar-se d'una patent, els detalls del mètode patentat no han estat publicats. La falta de solucions existents per a l'EQ de cada mot per a TAO basada en MT podria fer-vos pensar que aquesta tasca no és suficientment rellevant per despertar l'interés de la comunitat científica. Per refutar aquesta idea, l'Apèndix A de l'article reimprés 2.2.1 que es detalla més endavant descriu un experiment en el qual professionals de la traducció utilitzen una eina de TAO basada en MT per traduir textos de l'anglés a l'espanyol amb EQ per a cada mot i sense EQ per a cada mot. Aquests experiments confirmen que disposar d'EQ fiable pot reduir el temps dedicat a una tasca de traducció fins a un 14%. Aquest resultat confirma els avantatges que pot tenir per als traductors professionals aquesta tecnologia i, en conseqüència, emfatitza la rellevància de la recerca desenvolupada dins d'aquesta tesi doctoral en aquesta direcció.

El Capítol 2 presenta la tasca de l'EQ de la traducció per a TAO basada en MT. S'hi exploren dues vies per a obtenir aquestes estimacions, cadascuna en una secció: la Secció 2.1 descriu mètodes basats en alineaments de mots, mentre que la Secció 2.2 descriu mètodes basats en l'ús de FIB externes.

Els mètodes descrits en la Secció 2.1 utilitzen alineaments entre els mots en S i T per projectar la informació sobre els mots en S que concorden amb S' sobre els mots en T, a fi de proporcionar una EQ de cada mot. La figura 1 mostra un exemple d'aquesta tècnica en el qual el segment en anglés S' que cal traduir és *"the Asia-Pacific Association for Machine Translation"*, i el suggeriment de traducció anglés–català proporcionat és (S, T) is *("the European Association for Machine Translation", "l'Associació Europea per a la traducció automàtica")*. En la figura es destaquen els mots de S que concorden amb S' (en negreta), i es mostra com aquesta informació es projecta sobre T utilitzant els alineaments entre mots, representats per arestes que connecten S i T. Aquesta secció conté dos articles reimpresos:

- Esplà, M., Sánchez-Martínez, F. i Forcada, M.L. 2011. Using word alignments to assist computer-aided translation users by marking which target-side words to change or keep unedited. En *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, p. 81–89, 30–31 de maig de 2011, Lovaina, Bèlgica. [**Article reimprés 2.1.1**]
- Esplà-Gomis, M., Sánchez-Martínez, F. i Forcada, M.L. 2012. A simple approach to use bilingual information sources for word alignment. En *Procesamiento del Lenguaje Natural*, 49, p. 93–100. [**Article reimprés 2.1.2**]

L'article reimprés 2.1.1 descriu les tècniques desenvolupades per a l'EQ de cada mot basades en models estadístics d'alineament de mots (Och and Ney, 2003). Hom podria pensar que els models estadístics d'alineament de mots no poden ser considerats FIB segons la definició inclosa al principi d'aquest capítol. Tanmateix, aquesta és la tècnica més comunament usada per relacionar els mots entre dos segments en llengües diferents, un pas necessari per a estimar la qualitat de la traducció per a TAO basada en MT, tal com l'hem definida. Així doncs, l'article reimprés 2.1.1 té per objectiu confirmar la següent hipòtesi de treball:

Hipòtesi #1: és possible utilitzar alineaments de mots per a estimar la qualitat de la traducció per a TAO basada MT.

El treball desenvolupat amb models estadístics d'alineament de mots va posar els fonaments per a les etapes de recerca següents en les quals es van usar FIB. Els experiments descrits en l'article reimprés 2.1.1, en què s'avaluen diverses tasques de traducció entre l'anglés i l'espanyol, van proporcionar resultats prometedors i van mostrar que era possible estimar la qualitat de la traducció en eines TAO basades en MT amb una precisió i cobertura altes mitjançant models estadístics d'alineament de mots.

L'article reimprés 2.1.2 cerca una via per a convertir el mètode definit en l'article reimprés 2.1.1 en un mètode basat en l'ús de FIB externes. Per a fer-ho, proposa un nou mètode heurístic capaç d'alinear mots *al vol* fent servir FIB i que, per tant,

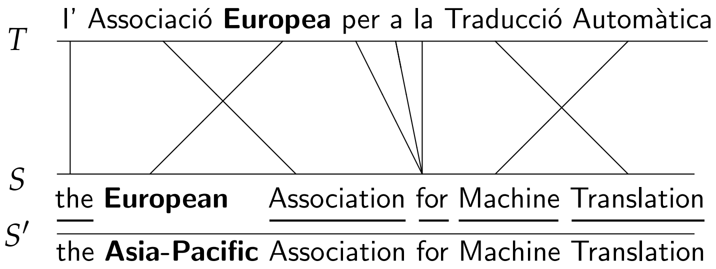

T
l' Associació **Europea** per a la Traducció Automàtica
S
the **European** Association for Machine Translation
S'
the **Asia-Pacific** Association for Machine Translation

Figura 1: Exemple d'EQ per a TAO basada en MT mitjançant l'alineament de mots. A l'exemple, els segments en LO S i S' són comparats, per trobar els mots en S que cal eliminar o reemplaçar (en negreta); després aquesta informació és projectada sobre T mitjançant l'alineament dels mots entre S i T.

elimina la dependència respecte dels models estadístics d'alineament de mots. Així doncs, la hipòtesi de treball que inspira aquest treball és:

Hipòtesi #2: és possible obtenir alineaments entre mots mitjançant l'ús de FIB.

El treball descrit en l'article reimprés 2.1.2 és ampliat a l'Apèndix A, on, a més, s'hi descriu un nou mètode més general que utilitza un model de *màxima versemblança*. Tant el mètode heurístic com el basat en el model de màxima versemblança són comparats amb l'eina més comunament usada per a l'alineament estadístic de mots: GIZA++ (Och and Ney, 2003). Els resultats obtinguts confirmen que els mètodes basats en FIB són capaços d'alinear mots amb una precisió comparable a l'obtinguda per GIZA++, tot i que, en general, la cobertura és més baixa. Els mètodes basats en FIB sols tenen una cobertura millor que GIZA++ quan els models estadístics d'alineament de mots són entrenats sobre un corpus paral·lel menut (al voltant de 10.000 parells de segments o menys).

Tot i que els resultats obtinguts amb alineament de mots basat en FIB no són tan acurats com s'esperava, aquest treball obri la porta a l'etapa següent de la recerca: l'EQ utilitzant FIB directament, la qual es descriu a la Secció 2.2 i conté una sola publicació:

- Esplà-Gomis, M., Sánchez-Martínez, F. i Forcada, M.L. 2015. Using machine translation to provide target-language edit hints in computer-aided translation based on translation memories. En *Journal of Artificial Intelligence Research*, volum 53, p. 169–222. [**Article reimprés 2.2.1**]

L'article reimprés 2.2.1 descriu dos mètodes diferents que fan servir FIB directament per a l'EQ de cada mot: un d'heurístic, i un que utilitza un classificador binari basat en aprenentatge automàtic. L'objectiu principal de la recerca descrita en aquest article és confirmar la següent hipòtesi de treball:

Hipòtesi #3: és possible utilitzar FIB directament per a estimar la qualitat en cada mot de la traducció en TAO basada en MT.

Els mètodes descrits a l'article reimprés 2.2.1 es comparen amb els mètodes basats en models estadístics d'alineament de mots proposats en l'article reimprés 2.1.1, per a cinc parells de llengües diferents: anglés–espanyol, anglés–francés, anglés–alemany, anglés–finés, i espanyol–francés. El marc d'avaluació proposat en aquest article és més fiable per als diferents mètodes descrits al Capítol 2, ja que aquests mètodes s'avaluen per a la traducció entre llengües molt properes (com ara l'espanyol i el francés, que són llengües romàniques, o l'anglés i l'alemany, que són llengües germàniques), entre llengües de la mateixa família, tot i no ser tan properes (l'anglés, l'espanyol, l'alemany i el francés són totes llengües indoeuropees, però les diferències entre les llengües germàniques i romàniques són substancials), i, fins i tot, entre llengües que no tenen cap relació entre elles (el finés és una llengua uràlica, i, per tant, no està relacionada de cap forma amb cap de les altres quatre llengües, que són indoeuropees). Els experiments descrits en aquest article confirmen que els resultats obtinguts amb els mètodes basats directament en FIB són en general millors que els obtinguts pels mètodes basats en models estadístics d'alineament de mots, especialment quan aquests han de traduir textos de dominis diferents als dels textos que s'han utilitzat per a entrenar els models d'alineament.

Ús de fonts d'informació bilingüe per a l'estimació de la qualitat de la traducció per a traducció automàtica

La segona tecnologia de la traducció en què aquesta tesi doctoral se centra és la TA. Trobem a la bibliografia diverses tècniques que aborden el problema de l'EQ en TA; la majoria, basades en aprenentatge automàtic. Aquestes tècniques basades en aprenentatge automàtic extrauen característiques de les traduccions mitjançant les quals és possible discernir quins mots són adequats i quins no ho són i, per tant, necessiten ser posteditats. Aquestes característiques es divideixen, principalment, en dues classes: les que necessiten accedir a les dades internes del sistema de TA que ha produït la traducció i les que són independents del sistema de TA (Quirk, 2004; Blatz et al., 2004; Specia et al., 2010). Tanmateix, fins a on sabem, totes les col·leccions de característiques disponibles a la bibliografia depenen d'una font d'informació específica, com ara models de llengua, lexicons bilingües, models de reordenament de mots, etc.; en altres paraules, cap d'aquestes col·leccions usa FIB d'una manera agnòstica. Per tant, l'objectiu d'aquest bloc de recerca és desenvolupar mètodes que, basant-se en els descrits al Capítol 2, siguen capaços d'estimar la qualitat de les traduccions produïdes per un sistema de TA utilitzant qualsevol FIB disponible.

Arribats a aquest punt, és important analitzar les diferències entre els problemes de l'EQ per a TA i per a TAO basada en MT: mentre en la TAO basada en MT el problema consisteix a detectar quins mots en una traducció adequada de S no són part de la traducció del nou segment S', en TA s'hi treballa sobre una traducció automàtica de S', la qual pot ser adequada o no. Per tant, tot i que l'objectiu és aprofitar els conceptes principals del mètode basat en FIB que s'ha desenvolupat per a l'EQ en TAO basada en MT, cal definir un mètode substancialment diferent per al cas de la TA.

Així, el Capítol 3 descriu un nou mètode basat en FIB que aborda el problema de l'EQ de cada mot per a TA amb un enfocament de classificació binària. Aquest mètode aplica la mateixa tècnica d'aprenentatge automàtic descrita en la Secció 2.2, però utilitzant noves característiques de les traduccions T, per marcar-ne els mots com a "bons" (no cal posteditar-los) o "roïns" (cal eliminar-los o substituir-los). En el cas de l'EQ per a TA, s'han definit dues famílies de característiques: una amb característiques positives, que proporcionen informació a favor que el mot siga marcat com a bo, i una altra amb característiques negatives, que indiquen que el mot podria haver de ser eliminat o substituït.

El Capítol 3 conté dues publicacions:

- Esplà-Gomis, M., Sánchez-Martínez, F. i Forcada, M.L. 2015. Using on-line available sources of bilingual information for word-level machine translation quality estimation. En *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, p. 19–26, Antalya, Turquia, 11–13 de maig de 2015. [**Article reimprés 3.1**]
- Esplà-Gomis, M., Sánchez-Martínez, F. i Forcada, M.L. 2015. UAlacant word-level machine translation quality estimation system at WMT 2015. En *Proceedings of the 10th Workshop on Statistical Machine Translation*, p. 309–315, Lisboa, Portugal, 17–18 de setembre de 2015. [**Article reimprés 3.2**]

Aquestes dues publicacions tenen com a objectiu confirmar la hipòtesi de treball següent:

> **Hipòtesi #4:** és possible adaptar les tècniques d'EQ desenvolupades per a TAO basada en MT al cas de la TA.

L'article reimprés 3.1 descriu el mètode basat en classificació binària proposat, així com les col·leccions de característiques que fa servir el classificador binari automàtic. A més, l'article conté una col·lecció d'experiments que serveixen per a avaluar el mètode proposat utilitzant les dades d'avaluació proporcionades pels organitzadors de la tasca compartida d'EQ de cada mot per a TA en l'edició de 2014 del *Workshop on Statistical Machine Translation* (Bojar et al., 2014).[3] Les dades usades en l'avaluació per a la tasca en 2014 eren disponibles per a dos parells d'idiomes: anglés–espanyol i anglés–alemany, en totes dues direccions de traducció. Tal com s'explica al Capítol 3, tot i les diferències entre les llengües, els resultats obtinguts no sols confirmen la viabilitat del mètode proposat, sinó que, a més, els sistemes desenvolupats en aquesta tesi demostren una gran eficàcia, comparable a la dels sistemes que van obtenir els millors resultats en aquesta edició de la tasca compartida.

L'article reimprés 3.2 descriu l'aplicació del mètode proposat en aquesta tesi a l'edició de 2015 de la tasca compartida d'EQ de cada mot per a TA (Bojar et al., 2015). En aquesta edició, les dades d'avaluació van ser proporcionades només per a la traducció de l'espanyol a l'anglés. A més, aquest any l'organització va proporcionar un conjunt de característiques bàsiques com a punt de partida per als sistemes desenvolupats. La combinació de les característiques definides a l'article reimprés 3.1 i les característiques bàsiques proporcionades per l'organització de la tasca van permetre al nostre sistema obtenir els millors resultats (Bojar et al., 2015) entre tots els participants de la tasca de 2015.

Construcció de noves fonts d'informació bilingüe per a parells de llengües amb pocs recursos

Un dels pilars principals d'aquesta tesi doctoral és la disponibilitat de FIB. De fet, tal com s'explica al principi d'aquesta introducció, un dels objectius de la recerca duta a terme és aprofitar la gran quantitat de FIB que són disponibles per al seu ús. Tanmateix, tal com podríeu haver pensat, aquesta suposició no és vàlida per a tots els parells de llengües. L'estudi de Rehm i Uszkoreit (2013), que té per objectiu analitzar les tecnologies lingüístiques disponibles per a 30 llengües europees (23 d'elles oficials a la Unió Europea), aporta dades que donen suport a aquesta idea. Una de les conclusions d'aquest informe és que "moltes llengües manquen fins i tot de les tecnologies bàsiques per a l'anàlisi de textos i de recursos lingüístics essencials".

Per mitigar la mancança de FIB per a alguns parells de llengües, part d'aquesta tesi doctoral s'ha centrat en desenvolupar un mètode per a crear noves FIB mitjançant

[3] La tasca compartida d'EQ de cada mot del *Workshop on Statistical Machine Translation* porta organitzant-se des de fa tres anys, i és un torneig en què s'avaluen sistemes d'EQ de cada mot desenvolupats pels concursants sobre unes dades d'avaluació comunes. Els organitzadors de la tasca proporcionen una col·lecció de segments en LO i les corresponents traduccions obtingudes amb un sistema de TA. Tres conjunts de dades són proporcionats: un d'entrenament, un de desenvolupament, i un de prova. Per als dos primers conjunts, els mots de les traduccions estan etiquetats com a "bons" i "roïns" (tot i que alguns anys també s'han proporcionat conjunts d'etiquetes amb un gra més fi per als diferents tipus d'errors de traducció), mentre que els participants han d'etiquetar els mots de les traduccions del conjunt de prova. L'ús d'un conjunt de dades comú proporciona un marc d'avaluació adequat per a comparar els sistemes desenvolupats per a la tasca, tal com ho són els que es descriuen en aquesta tesi doctoral.

l'ús de l'eina Bitextor (Esplà-Gomis i Forcada, 2010) (versió 4.1) per a la recol·lecció de textos paral·lels a partir de llocs webs multilingües. Aquesta eina descarrega llocs web multilingües i n'alinea els documents mitjançant: (i) l'ús de lexicons bilingües que permeten la comparació del contingut dels documents amb un mètode basat en el de Sánchez-Martínez i Carrasco (2011), i (ii) la comparació de l'estructura HTML dels documents (Resnik and Smith, 2003). A més, Bitextor és capaç d'alinear els documents per segments mitjançant l'eina Hunalign (Varga et al., 2005). Aquests corpus paral·lels alineats per segments poden ser fàcilment utilitzats per a construir noves FIB, com ara lexicons bilingües, taules de subsegments, o sistemes de TA estadística, que es poden usar amb les tècniques d'EQ de cada mot descrites als Capítols 2 i 3.

El Capítol 4 descriu la recerca duta a terme sobre la creació de noves FIB, i conté dues publicacions:

- Esplà-Gomis, M., Klubička, F., Ljubešić, N., Ortiz-Rojas, S., Papavassiliou, S. i Prokopidis, P. 2014. Comparing two acquisition systems for automatically building an English–Croatian parallel corpus from multilingual websites. En *Proceedings of the 9th International Conference on Language Resources and Evaluation*, p. 1252–1258, Reykjavík, Islàndia, 26–31 de maig de 2014. [**Article reimprés 4.1**]
- Toral, A., Rubino, R., Esplà-Gomis, M., Pirinen, T., Way, A. i Ramírez-Sánchez, G. 2014. Extrinsic evaluation of web-crawlers in machine translation: a case study on Croatian–English for the tourism domain. En *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, p. 221–224, Dubrovnik, Croàcia, 16–18 de juny de 2014. [**Article reimprés 4.2**]

Tal com s'indica al prefaci d'aquesta tesi doctoral, la major part de la recerca descrita en aquest capítol s'ha desenvolupat en el marc del projecte Abu-MaTran,[4] finançat per la Unió Europea, el qual se centra en les llengües eslaves del sud, parant una especial atenció al croat. Per aquest motiu, tots dos articles se centren en la creació de FIB per al parell de llengües anglés–croat. La recerca que s'hi descriu té com a objectiu confirmar la hipòtesi de treball següent:

Hipòtesi #5: és possible crear noves FIB per a l'EQ de cada mot per a parells de llengües sense cap FIB disponible utilitzar Bitextor per a recol·lectar corpus paral·lels.

L'article reimprés 4.1 descriu l'avaluació intrínseca del corpus paral·lel anglés–croat recol·lectat a partir de 21 llocs web amb Bitextor i un altre sistema actual per a la recol·lecció de textos paral·lels: l'ILSP Focused Crawler (Papavassiliou et al., 2013). L'article descriu els resultats obtinguts per totes dues eines, en termes de

[4] `http://www.abumatran.eu`

quantitat de text paral·lel obtingut i qualitat del corpus construït, i els compara.[5] Els resultats prometedors que es van obtenir en aquests experiments, especialment pel que fa a la qualitat dels corpus, van motivar la recerca descrita a l'article reimprés 4.2, on s'avaluen aquests corpus extrínsecament. Per a fer-ho, tots dos corpus van ser utilitzats per a entrenar un sistema de TA estadística basat en sintagmes (Koehn et al., 2003), que va ser avaluat en una tasca de traducció entre l'anglés i el croat. Els resultats d'aquest article confirmen la utilitat de les dades recol·lectades per a la creació d'un sistema de TA estadística plenament funcional.

Tot i els bons resultats descrits als articles reimpresos 4.1 i 4.2, en el moment de publicar aquesta memòria encara no s'havia publicat una avaluació de l'impacte de FIB creades amb Bitextor per a l'EQ. Per aquest motiu, l'Apèndix B informa sobre els resultats dels experiments addicionals duts a terme en aquest sentit, l'objectiu dels quals és confirmar la darrera hipòtesi de treball d'aquesta tesi doctoral:

Hipòtesi #6: els resultats obtinguts per a l'EQ de cada mot per a parells de llengües amb pocs recursos poden ser millorats mitjançant l'ús de noves FIB obtingudes a través de la recol·lecció de corpus paral·lels.

Els nous experiments duts a terme recuperen alguns dels experiments descrits a la Secció 2.2 i se centren en el parell de llengües amb menys recursos d'aquells descrits en la Secció 6 de l'article reimprés 2.2.1: l'anglés–finés. Els experiments originals mostraven que, a causa de la cobertura relativament reduïda de les FIB disponibles per a aquest parell de llengües, la qualitat d'una part important de els mots al conjunt de prova (més del 10%) no havia pogut ser estimada. La baixa cobertura de FIB feia que no es poguera projectar la informació sobre els mots de S que concordaven amb S' sobre T. Els experiments descrits a l'Apèndix B.2 demostren que la quantitat de mots per als quals no es pot estimar la qualitat cau dramàticament quan s'utilitzen les FIB obtingudes amb Bitextor.

Discussió

En conclusió, en aquesta tesi doctoral s'han descrit un seguit de mètodes que permeten l'EQ de cada mot per a dues tecnologies, la TAO basada en MT i la TA, fent servir FIB. L'objectiu principal d'aquests mètodes és el d'aprofitar les FIB que són disponibles, per exemple a Internet, i donar-los un nou ús en l'àmbit de la traducció.

En aquesta tesi es descriuen, per primera vegada, mètodes que permeten l'EQ de cada mot per a TAO basada en MT. La rellevància d'aquesta tasca ha estat avaluada mitjançant la realització d'experiments amb traductors professionals, i s'ha demostrat que l'EQ de cada mot en TAO pot permetre estalviar fins al 14% del temps

[5] Una fracció aleatòria dels corpus va ser analitzada manualment per estimar-ne la qualitat.

invertit en una tasca de traducció. Els mètodes desenvolupats han estat avaluats en múltiples tasques de traducció amb diferents condicions, com ara les llengües a traduir, el domini dels textos o les FIB utilitzades. En tots els casos, la viabilitat dels mètodes ha estat demostrada.

Els mètodes per a l'EQ de cada mot en TAO basada en MT han estat, posteriorment, ampliats a una segona tecnologia de la traducció: la TA. En el cas d'aquesta tecnologia, la bibliografia conté nombrosos treballs sobre EQ. Tanmateix, la idea d'usar FIB com a font d'informació és nova, ja que la resta de mètodes desenvolupats depenen de fonts d'informació específiques. Més enllà de l'originalitat en l'ús de FIB per a l'EQ, l'avaluació mitjançant les dades de les tasques compartides en EQ de cada mot per a TA en les edicions de 2014 i 2015 del *Workshop on Statistical Machine Translation* ha demostrat que els resultats obtinguts pels mètodes desenvolupats en aquesta tesi doctoral són comparables als sistemes més reeixits en aquesta tasca. Podem concloure, per tant, que l'ús de FIB no només permet reaprofitar recursos ja disponibles per a l'EQ, sinó que, a més, permet assolir les màximes quotes d'eficàcia en aquesta tasca.

Finalment, i com a complement de la recerca desenvolupada en el camp de l'EQ, cal destacar els resultats obtinguts pel que fa a la creació de noves FIB per al seu ús en EQ. La darrera part d'aquesta tesi doctoral s'ha enfocat a l'estudi de l'impacte que les FIB obtingudes automàticament mitjançant la recol·lecció de dades paral·leles a partir d'Internet poden tenir en aquesta tasca. Així, d'una banda, s'ha estudiat l'ús de l'eina Bitextor en la creació de FIB per a un parell de llengües amb pocs recursos: anglés–croat. Aquesta recerca ha demostrat una gran eficiència de l'eina a l'hora de crear corpus paral·lels, tant pel que fa a la quantitat de dades obtingudes, com a la qualitat d'aquestes. També s'ha estudiat l'ús dels corpus paral·lels recol·lectats per a la creació de sistemes de traducció automàtica, amb resultats molt positius. Finalment, l'ús de les FIB obtingudes amb Bitextor ha estat avaluat per al cas de l'EQ de cada mot per a la TAO basada en MT. Els experiments duts a terme han confirmat que l'ús de noves FIB creades expressament per a aquesta tasca en millora el rendiment dramàticament, especialment quan es tracta de llengües amb pocs recursos.

Un dels elements clau d'aquesta tesi doctoral és que defineix, per primera vegada, una estratègia per a l'EQ que utilitza les mateixes fonts d'informació tant per a la TAO basada en MT com per a la TA. Açò significa que aquestes estratègies podrien ser integrades en un sistema de TAO que implemente totes dues tecnologies de traducció per a estimar la qualitat dels suggeriments de traducció provinents d'ambdues fonts en paral·lel i mitjançant les mateixes FIB. Així, els traductors podrien gaudir del suport d'aquestes tècniques sense haver de crear models específics per a cadascuna d'elles. Fins i tot seria possible integrar l'eina Bitextor en aquest entorn de TAO per proporcionar suport a aquells parells de llengües per als quals l'usuari no disposara de FIB, permetent la màxima disponibilitat de l'EQ dins d'aquest entorn.

Preface

Translation technologies are aimed at aiding professional translators in the translation process so that they are more productive. In fact, some of these technologies are able to perform the bulk of a translation, reducing the task of professional translators to review and, if necessary, post-edit the translations to make them adequate for the intended purpose.

The evolution of translation technologies has led the scientific community to focus on the development of methods capable of providing translators with an estimation of the quality of the output produced by a specific translation technology. Machine translation has captured most of the attention as regards quality estimation during the last years, since it is the only translation technology capable of producing complete translations (at the level of segments or paragraphs) in a fully automatic way. However, other translation technologies have been using quality estimation indicators since their inception, this is the case of computer aided translation tools based on translation memories, which use fuzzy-match scores to inform the translator about the *usefulness* of translation suggestions.

Aiding translators to determine the quality of a translation, regardless of the translation technology producing it, may be of great help because it allows translators to make decisions even before they start translating or post-editing; decisions such as which technology is more helpful for a translation task, or how much effort a translation task may require. Estimating the quality of translations is a complex problem that may require data that is not always available, such as information about the inner workings of the system used to produce the translations, or complex models that require large amounts of monolingual or bilingual data to be trained.

The work described in this dissertation is aimed at developing methods to estimate the quality of the translations produced by different translation technologies by using external sources of bilingual information. The initial focus was on methods for assisting users of computer-aided translation tools based on translation memories. These tools allow reusing previous translations stored in a translation memory for a new translation task by automatically looking for those translation units with a source language segment *similar*[6] to that to be translated. These similar translation units are provided to translators so that they can use their target language segments as a basis for the new translation, avoiding the need to translate the source-language segment from scratch. Initially, the objective was to design methods able to detect the words that need to be post-edited in the translation suggestions provided by these tools. However, the growing knowledge acquired on this field as the research went on highlighted the obvious parallelisms between this problem and the problem known as *word-level quality estimation* in machine translation. Widening the scope to include machine translation and extending the objective of this dissertation to other translation technologies became a natural choice at some point.

Given that most of the research conducted as part of this PhD thesis has been published in international peer-reviewed conferences and journals, this dissertation is configured as a compilation of papers, which means that each chapter will contain the papers published in each of the corresponding research blocks. These papers are rendered as sections, even though they are reprinted keeping their original format; throughout this dissertation they will be referred by their section number.

Part of the research described in this dissertation has been developed in the framework of the Abu-MaTran EU-funded project, which is aimed at developing methods for automatically building machine translation systems. This project devotes especial attention to machine translation between South-Slavic languages and English; this was motivated by the scarce bilingual resources available for these language pairs. My contribution to this project has mainly focused on the task of crawling parallel data by adapting the tool Bitextor (Esplà-Gomis and Forcada, 2010) for the new languages involved in the project. I started working on this tool as part of my bachelor thesis in 2008, and I kept working on it during my master's studies in 2010. This dissertation contains a chapter devoted to the discussion of the usefulness of crawling parallel data to create new sources of bilingual information (not only machine translation, but also bilingual dictionaries, phrase tables, etc.) for under-resourced language pairs and their application to word-level quality estimation. Unfortunately, given the bureaucratic restrictions at Universitat d'Alacant, it was impossible to include the publications regarding my work on Bitextor because they were published before I "officially" started my PhD studies. Despite this, they should be considered as part of my PhD work.

This dissertation is organised in five chapters: Chapter 1 provides an introduction that, among other things, puts every chapter in context; chapters 2–4 consist of papers already published; and Chapter 5 provides some concluding remarks and describes open research lines. Additionally, two appendices are included in this dissertation that describe additional methods and experiments: one containing a

[6] A variety of similarity metrics are used by computer-aided translation tools based on translation memories; most of them are based on the edit distance.

technical report on methods for word alignment based on sources of bilingual information, and another containing additional experiments aimed at evaluating: (i) the use of the word-alignment methods based on sources of bilingual information for word-level quality estimation in computer-aided translation, and (ii) the impact of using sources of bilingual information obtained with Bitextor for word-level quality estimation in computer-aided translation for under-resourced language pairs. The reason for having these appendices is that the research work they describe has not been published in peer-reviewed conferences or journals and their addition is necessary for a complete reporting of the research conducted.

This thesis has been possible thanks to the ideas and constant supervision of Dr. Felipe Sánchez-Martínez and Dr. Mikel L. Forcada from the Departament de Llenguatges i Sistemes Informàtics at Universitat d'Alacant. The work and ideas by the researchers collaborating in the Abu-MaTran project have been particularly valuable for the research conducted regarding the crawling of parallel data and the creation of new sources of bilingual information.

Structure of this dissertation

This dissertation is structured in 5 chapters and 3 appendices:

Chapter 1 introduces the most important concepts and definitions about quality estimation in translation technologies and explains the motivation behind the development of the new approaches described in this dissertation. It also puts the publications provided in chapters 2 to 4 into the context of my PhD work.

Chapter 2 presents the idea of word-level quality estimation in computer-aided translation based on translation memories and describes the methods proposed to tackle this problem and the results achieved.

Chapter 3 explains a novel approach using external sources of bilingual information for word-level machine translation quality estimation, and compares it to other state-of-the-art approaches.

Chapter 4 describes the research conducted on the creation of new sources of bilingual information from multilingual websites for under-resourced language pairs.

Chapter 5 summarises the main contributions to the state of the art of the work reported in this dissertation and outlines some future research lines.

Appendix A describes a maximum-likelihood-style model to obtain word alignments from sources of bilingual information; this model extends a heuristic method previously described in Chapter 2.

Appendix B presents a collection of additional experiments concerning the techniques described in chapters 2 and 4: experiments on the application of word-alignment methods based on sources of bilingual information for word-level quality estimation, and experiments on using new sources of bilingual information obtained from multilingual websites to improve the performance of word-level quality estimation. In both cases, the evaluation is performed in the context of computer aided translation based on translation memories.

Appendix C describes the software developed to evaluate the novel approaches proposed in this PhD thesis, and links each software package with the experiments conducted in each chapter.

Publications

Most of the content of this dissertation has been published in journals and peer-reviewed conference or workshop proceedings. The list below shows them in inverse chronological order. The chapter in which each publication appears is shown in brackets:

- Esplà-Gomis, M., Sánchez-Martínez, F., and Forcada, M.L. 2015. UAlacant word-level machine translation quality estimation system at WMT 2015. In *Proceedings of the 10th Workshop on Statistical Machine Translation*, p. 309–315, Lisbon, Portugal, September 17–18, 2015. **[Chapter 3]**

- Esplà-Gomis, M., Sánchez-Martínez, F., and Forcada, M.L. 2015. Using machine translation to provide target-language edit hints in computer aided translation based on translation memories. In *Journal of Artificial Intelligence Research*, volume 53, p. 169–222. **[Chapter 2]**

- Esplà-Gomis, M., Sánchez-Martínez, F., and Forcada, M.L. 2015. Using on-line available sources of bilingual information for word-level machine translation quality estimation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, p. 19–26, Antalya, Turkey, May 11–13, 2015. **[Chapter 3]**

- Toral, A., Rubino, R., Esplà-Gomis, M., Pirinen, T., Way, A., and Ramírez-Sanchez, G. 2014. Extrinsic evaluation of web-crawlers in machine translation: a case study on Croatian–English for the tourism domain. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, p. 221–224, Dubrovnik, Croatia, June 16–18, 2014. **[Chapter 4]**

- Esplà-Gomis, M., Klubička, F., Ljubešić, N., Ortiz-Rojas, S., Papavassiliou, S., and Prokopidis, P. 2014. Comparing two acquisition systems for automatically building an English–Croatian parallel corpus from multilingual websites. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, p. 1252–1258, Reykjavík, Iceland, May 26–31, 2014. **[Chapter 4]**

- Esplà-Gomis, M., Sánchez-Martínez, F., Forcada, M.L. 2012. A simple approach to use bilingual information sources for word alignment. In *Procesamiento del Lenguaje Natural*, 49, p. 93–100. **[Chapter 2]**

- Esplà, M., Sánchez-Martínez, F., Forcada, M.L. 2011. Using word alignments to assist computer-aided translation users by marking which target-side words to change or keep unedited. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, p. 81–89, May 30–31, 2011, Leuven, Belgium. **[Chapter 2]**

Some other papers related to my PhD work have been published in peer-reviewed conference or workshop proceedings but have not been included in this dissertation because the work they report is basically contained in the reprinted papers already included. The list below shows them in inverse chronological order, again including the chapter to which they are related to:

- Rubino, R., Pirinen, T., Esplà-Gomis, M., Ljubešić, N., Ortiz Rojas, S., Papavassiliou, V., Prokopidis, P., and Toral, A. 2015. Abu-MaTran at WMT 2015 translation task: morphological segmentation and web crawling. In *Proceedings of the 10th Workshop on Statistical Machine Translation*, p. 184–191, Lisbn, Portugal, September 17–18, 2015. **[Chapter 4]**

- Esplà-Gomis, M., Sánchez-Martínez, F., Forcada, M.L. 2011. Using machine translation in computer-aided translation to suggest the target-side words to change. In *Proceedings of the 13th Machine Translation Summit*, p. 172–179, Xiamen, China, September 19–23, 2011. **[Chapter 2]**

I have also published other papers in peer-reviewed conference proceedings that, although not directly related to the work presented in this dissertation, are linked to the creation of new sources of bilingual information and translation quality estimation.

- Ljubešić, N., Esplà-Gomis, M., Klubička, F., and Preradović, N.M. 2015. Predicting inflectional paradigms and lemmata of unknown words for semi-automatic expansion of morphological lexicons. In *Proceedings of Recent Advances in Natural Language Processing*, p. 379–387, Hissar, Bulgaria, September 5–11, 2015.

- Esplà-Gomis, M., Sánchez-Cartagena, V.M., Sánchez-Martínez, F., Carrasco, R.C., Forcada, M.L., and Pérez-Ortiz, J.A. 2014. An efficient method to assist non-expert users in extending dictionaries by assigning stems and inflectional paradigms to unknown words. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, p. 19–26, Dubrovnik, Croatia, June 16–18, 2014.

- C. de Souza, J.G., Esplà-Gomis, M., Turchi, M., and Negri, M. 2013. Exploiting qualitative information from automatic word alignment for cross-lingual NLP tasks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 771–776, Sofia, Bulgaria, August 4–9, 2013.

- Sánchez-Cartagena, V.M., Esplà-Gomis, M., Sánchez-Martínez, F., Pérez-Ortiz, J.A. 2012. Choosing the correct paradigm for unknown words in rule-based machine translation systems. In *Proceedings of the 3rd Workshop on Free/Open-Source Rule-Based Machine Translation*, p. 27–37, Gothenburg, Sweeden, June 14–15, 2012.

- Esplà-Gomis, M., Sánchez-Martínez, F., Forcada, M.L. 2012. UAlacant: Using online machine translation for cross-lingual textual entailment. In *Proceedings of the *SEM 2012: The 1st Joint Conference on Lexical and Computational Semantics (SemEval task 8)*, p. 472–476, Montreal, Quebec, Canada, June 7–8, 2012.

- Sánchez-Cartagena, V.M., Esplà-Gomis, M., Pérez-Ortiz, J.A. 2012. Source-language dictionaries help non-expert users to enlarge target-language dictionaries for machine translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, p. 3422–4429, Istanbul, Turky, May 23–25, 2012.

- Esplà-Gomis, M., Sánchez-Cartagena, V.M., Pérez-Ortiz, J.A. 2011. Multimodal building of monolingual dictionaries for machine translation by non-expert users. In *Proceedings of the 13th Machine Translation Summit*, p. 147–154, Xiamen, China, September 19–23, 2011.

- Esplà-Gomis, M., Sánchez-Cartagena, V.M., Pérez-Ortiz, J.A. 2011. Enlarging monolingual dictionaries for machine translation with active learning and non-expert users. In *Proceedings of Recent Advances in Natural Language Processing*, p. 339–346, Hissar, Bulgaria, September 12–14, 2011.

Contents

Preface **xix**

1 Introduction **1**

1.1 Translation quality in translation technologies 1

1.1.1 Quality estimation for machine translation 3

1.1.2 Quality estimation for computer-aided translation based on translation memories 5

1.2 Creating new sources of bilingual information for under-resourced language pairs 6

1.3 Problems addressed 8

1.3.1 Using sources of bilingual information for word-level quality estimation in computer-aided translation based on translation memories 10

1.3.2 Using sources of bilingual information for word-level quality estimation in machine translation 13

1.3.3 Building new sources of bilingual information for under-resourced language pairs by crawling parallel data 15

2 Quality estimation in computer-aided translation based on translation memories **17**

2.1 Methods for word-level quality estimation in computer-aided translation based on translation memories using word alignments 17

2.1.1 **Reprinted publication**: Using word alignments to assist computer-aided translation users by marking which target-side words to change or keep unedited 21

2.1.2 **Reprinted publication**: A simple approach to use bilingual information sources for word alignment 29

2.2 Methods for word-level quality estimation in computer-aided translation based on translation memories using external sources of bilingual information 36

2.2.1 **Reprinted publication**: Using machine translation to provide target-language edit hints in computer aided translation based on translation memories 39

3 Methods for word-level quality estimation in machine translation using external sources of bilingual information 93

3.1 **Reprinted publication**: Using on-line available sources of bilingual information for word-level machine translation quality estimation . . 97

3.2 **Reprinted publication**: UAlacant word-level machine translation quality estimation system at WMT 2015 105

4 Creation of sources of bilingual information for under-resourced language pairs from multilingual websites 113

4.1 **Reprinted publication**: Comparing two acquisition systems for automatically building an English–Croatian parallel corpus from multilingual websites 117

4.2 **Reprinted publication**: Extrinsic evaluation of web-crawlers in machine translation: a case study on Croatian–English for the tourism domain 125

5 Concluding remarks 129

5.1 Summary 129

5.2 Future work 133

A Extended methods for word alignment using external sources of bilingual information 137

A.1 **Reprinted publication**: Using external sources of bilingual information for on-the-fly word alignment 139

B Additional experiments 159

B.1 Using word-alignment methods based on sources of bilingual information for word-level quality estimation in computer-aided translation160

B.2 Using Bitextor to build new sources of bilingual information for word-level quality estimation in computer-aided translation 164

C Open-source software released 171

C.1 `Gamblr-CAT` 171

C.2 `OmegaT-Marker-Plugin` 172

C.3 `OmegaT-SessionLog` 173

C.4 `Flyligner` 173

C.5 `Gamblr-MT` 174

C.6 `Bitextor v.4.1` 174

Index of abbreviations **177**

Bibliography **179**

Chapter 1

Introduction

Translation quality estimation (QE) is the task of predicting the level of quality of the translation provided for a given source language (SL) text. QE has become crucial for translation technologies: translators need accurate QE to estimate the effort needed for a translation task and to choose the translation technologies to be used. This dissertation describes a collection of new techniques for word-level QE for two translation technologies: machine translation and computer-aided translation (CAT) based on translation memories (TM).

The methods for word-level QE proposed in this dissertation use any source of bilingual information (SBI) in an agnostic fashion, that is, without making any assumptions regarding the amount, quality, or format of the bilingual information used. In the context of this PhD thesis, we will refer as SBI to any resource able to provide a translation in one language for a given sub-segment in a different language. For the methods developed to be applicable to under-resourced language pairs, research has also been conducted on building new SBI by means of the acquisition of resources from the web.

The aim of this chapter is to introduce the basic concepts of QE for translation technologies, to present the motivation of the research developed in the framework of this dissertation, and to put the reprinted papers included in the rest of this dissertation in a common context.

1.1 Translation quality in translation technologies

The concept of translation quality is rather complex itself. House (1997) points out that there exist as many definitions of translation quality as theories of translation. Fields et al. (2014), Melby et al. (2014), and Koby et al. (2014), try to shed some light on this concept through their trilogy of articles, in which they analyse, respectively: (i) the current landscape of translation, (ii) the meaning of "quality" in an industrial environment, and (iii) the concept of translation quality. Even though several definitions of translation quality are proposed, the authors acknowledge that they "strongly disagree about which definition of translation quality is better for the translation industry" (Koby et al., 2014).

One may think that evaluating the quality produced by a translation tool would require a more specific definition of quality that allows to assess the level of correctness of a translation from an empirical point of view. However, most of the strategies developed for the evaluation of translation technologies simply rely on the comparison to human (reference) translations instead of proposing a definition of quality. We therefore find two options for translation quality evaluation in the literature: (i) using human judgements on translation hypotheses, or (ii) using reference translations (produced by humans) to automatically evaluate translation hypotheses. In the first case, translation hypotheses can be evaluated in general, as good or bad, or they can be evaluated for different features independently, usually considering separately their accuracy and fluency (Koehn, 2010, Chapter 8.1). Automatic evaluation based on reference translations is proposed as an alternative that allows to systematise the way to evaluate and, therefore, to reduce the subjectivity of human evaluators.[1]

There are mainly two kinds of automatic metrics for translation quality evaluation: those that measure the quality of a translation and those that measure the amount of errors in a translation. Among the metrics measuring the quality, BLEU (Papineni et al., 2002), NIST (Doddington, 2002) and METEOR (Banerjee and Lavie, 2005) are the most relevant metrics, where higher scores indicate higher translation quality. These metrics are based on the fraction of n-grams[2] that appear both in the translation hypothesis and the reference translations used for evaluation. On the other hand, word-error rate (WER) and translation edit rate (TER) (Olive, 2005) are two of the most popular metrics for translation quality evaluation measuring the amount of errors in a translation. To obtain the number of translation errors, both metrics use the edit distance (Levenshtein, 1966) to a reference translation. In contrast to the quality metrics, for error metrics lower scores mean higher translation quality. TER and WER are quite similar; the first one counts the number of editions (insertions, deletions, or replacements) at the word level that need to be carried out in order to transform the translation hypothesis into the reference translation, while the second metric works similarly, but introducing a new edit operation: block movement. A modified version of TER, human-targeted TER (HTER), was proposed by Snover et al. (2006) to include humans in the loop of automatic evaluation. HTER

[1] It is worth noting that this subjectivity may remain insofar as many valid translations may be produced for a sentence and the ones used for evaluation are still produced by humans.

[2] METEOR uses only uni-grams, while NIST and BLEU use n-grams usually in the range $[1, 4]$.

uses the TER metric on a reference translation created by a human translator by post-editing the very same translation hypothesis to be evaluated. This method reduces the subjectivity introduced by the translator when creating a reference translation and, therefore, makes HTER a more reliable metric, even though it is no longer automatic.

Despite the advances in translation quality evaluation, all the methods developed to date share the same problem: they depend on reference translations produced by humans. This means that they are useful for evaluating a given translation system, but they have limited use when it comes to assist a user when producing new translations with it. To cope with this situation, translation QE is proposed as the problem of predicting the quality of a translation without the need for reference translations (Specia et al., 2010); a unified framework to integrate both translation evaluation and translation QE has recently been proposed by Forcada and Sánchez-Martínez (2015).

During the last years, considerable scientific effort has been devoted to QE, especially in the case of machine translation (MT), because of the functionalities that accurate translation QE may provide: estimation of the post-editing effort needed for a translation, more reliable budgeting for a translation task, and selection of the best translation among a collection of translation hypotheses. Word-level QE of translated text has an additional advantage: it may be used to guide the post-editor to choose the words to be modified in a translation hypothesis. Having reliable word-level QE would have a positive impact in the productivity of post-editors, but it may also open the door to new technologies that could even be more helpful, such as the automatic selection of the translation technology to be used for a translation task (Forcada and Sánchez-Martínez, 2015), or the automatic post-editing of translation hypotheses (Chander, 1998; Allen and Hogan, 2000; Béchara et al., 2012; Kuhn et al., 2010).

1.1.1 Quality estimation for machine translation

Since its inception, MT of a quality comparable to that produced by professional translators has been the ″holy grail″, both for the industry using this technology and the scientific community developing it. As for many other technologies, the inflated expectations in MT in its early days ended up disappointing institutions and companies investing on it. The most noticeable sign of this disappointment is the well-known report published in 1966 by the *Automatic Language Processing Advisory Committee* (ALPAC) established by the US government in 1964. In this report, it was stated that MT "presumably means going by algorithm from machine-readable source text to useful target text, without recourse to human translation or editing". According to this definition the conclusion of the report was clear: "in this context, there has been no MT of general scientific text, and none is in immediate prospect".

Today, almost 50 years after the ALPAC report, despite the obvious improvement of MT, we are still far from obtaining translations interchangeable with those produced by professional translators. However, both industry and academia are more realistic about the advantages and limitations of translation technologies. MT has been shown to be helpful as an *assimilation* or *gisting* tool (Hutchins, 2001) to access information in languages not understood by a reader, especially when the reader does not need to understand all the details of the text with precision. MT has also proved to be useful for *dissemination* of translated information; in most such cases, post-editing and careful revision of the translated text is required: it has been proved that, for some applications and language pairs, translating a text by using MT and then post-editing reduces the cost of professional translation (Plitt and Masselot, 2010) when compared to translating the text from scratch. In fact, despite the negative prospects for human-like quality of MT given, the ALPAC report also stated that "machine-aided translation may be an important avenue toward better, quicker, and cheaper translation". One can safely say that it has been in a CAT context where MT for dissemination purposes has found its place in industry.

Many technologies have been developed around MT to make it more usable, especially as a CAT component. This PhD thesis focuses on one of them: QE of translation, and, more specifically, QE at the word level. Some of the earliest approaches to MT QE can be found in the context of interactive MT (Gandrabur and Foster, 2003; Ueffing and Ney, 2005). Given a segment to be translated, interactive MT (IMT) systems use MT to suggest one or more translation options. The user of the tool may accept one of these suggestions or just translate by hand the next words; in any case, as the translation progresses, new translation suggestions are produced until the whole segment is translated. In this scenario, Gandrabur and Foster (2003) define a method to obtain confidence scores for the words in each translation suggestion in order to help the system to choose the most promising translation suggestions. Ueffing and Ney (2005) are the authors of one of the first word-level QE proposals, extending the approach by Gandrabur and Foster (2003) to be able, not only to choose among several translation suggestions, but also to be able to determine which words in these suggestions should be deleted or replaced.

Blatz et al. (2003) were among the first to propose a machine-learning approach for QE. They defined an extensive heterogeneous collection of features that can be extracted from the output of an MT system and used by a machine-learning model to estimate the quality of the translation, either at the segment level or the word level. A deeper study of those features that are independent of the MT system used to produce the translation hypothesis was performed by Specia et al. (2010). The authors show that using these features it is possible to estimate the quality of a translation produced by any MT system. One of the most complete collections of features in MT QE has been defined in the framework of the QuEst project (Specia et al., 2013). In QuEst, the features for QE are divided into three groups: those measuring the complexity of the SL segment S, those measuring the confidence on the MT system, and those measuring both fluency and adequacy directly on the translation hypothesis T. Many approaches have been proposed using such features. Some of them try to learn judgements from a gold-standard consisting of a collection of MT outputs and a confidence score, either produced by an automatic evaluation metric or a human evaluator (Quirk, 2004; Blatz et al., 2004). A different approach is proposed by Gamon et al. (2005), who try to learn to identify when a translation produced with MT *looks like* a human translation, by using a collection of human translations and MT outputs. Other authors base their approaches on the use of pseudo-references (Albrecht and Hwa, 2007; Biçici, 2013), that is, translations that are expected to be very similar to a reference translation, such as the previous translation of a very similar SL text, or a synthetic translation produced automatically by translating the SL text using a different MT system. Pseudo-references are used to infer quality information about the translation hypothesis by comparing them. It is worth noting that all of the approaches in the literature depend on specific sources of bilingual or monolingual information. Even those methods that do not access the inner workings of the MT system that produced the translation to be evaluated are strongly dependent on specific sources of information. For example, some of the features defined in the QuEst framework (Specia et al., 2013) depend on language models, statistical word alignment models, or probabilistic lexicons. Something similar happens with the approaches based on pseudo-references: these approaches need to train models capable to provide these pseudo-references and, for that, they require large amounts of parallel and monolingual data (Biçici, 2013). However, none of the approaches proposed until now are able to take advantage of the SBI already available in a robust and agnostic way, that is, without making any assumptions regarding the amount, quality, or format of the bilingual information used.

1.1.2 Quality estimation for computer-aided translation based on translation memories

Despite the effort devoted to develop methods in MT QE, this has not happened for the other technology on which this dissertation is focused: CAT based on translation memories (TM). As mentioned above, TM-based CAT tools help the translator recycle previous translations of segments stored in a TM for a new translation task. Although it is a rather simple tool, it is in fact one of the most popular translation technologies among professional translators (Bowker, 2002; Somers, 2003). In fact, one of the key points of the success of this technology is that it already provides a sentence-level QE metric: the fuzzy-match score (Sikes, 2007). Fuzzy-match scores (FMS) are usually based on the edit distance (Levenshtein, 1966) and compare the SL segment S' to be translated and the SL segments in the translation units (TU)

$\{(S, T)\}$ stored in the TM.[3] The basic idea behind TM-based CAT tools is that this comparison between the SL segments S' and S provides information to estimate the usefulness of the target language (TL) segment T of each TU and give useful information that the translator can use to edit T. FMS take values in $[0\%, 100\%]$, where 100% represents a translation hypothesis for S' that would not require any edition, and 0% a translation hypothesis which does not offer any help. TM-based CAT tools usually allow to set a FMS threshold to stop the system from suggesting translations with a very low FMS. This threshold is usually set by translators to values above 60% (Bowker, 2002, p. 100).

TM-based CAT tools provide segment-level QE through the FMSs, but basically no word-level QE. On the one hand, most TM-based CAT tools highlight the words in the SL segment S of a TU (S, T) suggested to the user that differ from those in the SL segment to be translated S'. However, no word-level QE assessment is provided for the TL segment in the TUs suggested. Most of the work in this direction has focused on automatically post-edit or "repair" the TL segments suggested by the TM-based CAT tools (Kranias and Samiotou, 2004; Ortega et al., 2014). For example, Kranias and Samiotou (2004) align the words in each TU (S, T) to determine which words in T need to be post-edited. However, the user is not informed about the results of this process; instead of this, the system defined by Kranias and Samiotou (2004) directly post-edits these words by using partial translations from MT. A similar approach is followed by Ortega et al. (2014), who use MT to produce translation "patches" consisting of one or more words that are applied, where they match, to the TL segment T in the TU suggested by the CAT tool, to create a "repaired" translation suggestion.

1.2 Creating new sources of bilingual information for under-resourced language pairs

Most of the work discussed in this dissertation relies on the assumption that a large amount of SBI are ready to use, for example, in the Internet. However, this is not the case for all language pairs. According to the study by Rehm and Uszkoreit (2013), aimed at analysing the support of language technologies for 30 European languages (23 of them being official languages of the European Union), "many languages lack even basic technologies for text analytics and essential language resources". Given that most of the methods for word-level QE described in this dissertation depend on the availability of SBI, the fact that many languages have such a poor technological support becomes a problem. This motivates the last block of research in this PhD

[3] Some commercial tools have also introduced sub-segment level fuzzy-match methods; for example, the tool *Déjà Vu* integrates example-based MT in order to produce suggestions by combining partial matches (Lagoudaki, 2008)

thesis: the development of methods to build new SBI for under-resourced language pairs. One of the most successful strategies targeting this objective during the last years has been to crawl parallel data from multilingual websites (Resnik and Smith, 2003; Sridhar et al., 2011; San Vicente and Manterola, 2012).

Large amounts of parallel data have been produced through automatic alignment of segments using well structured sources of translations. Some well known examples are multilingual parallel corpora built on documentation from official institutions, such as the Hansards corpus (Roukos et al., 1995), an English–French corpus crawled from the official records of the Canadian Parliament, the Europarl corpus (Koehn, 2005), a multilingual corpus with 21 European languages crawled from the proceedings of the European Parliament, or the MultiUN corpus (Eisele and Chen, 2010), a parallel corpus with 7 languages crawled from the official documents of the United Nations. Another source of parallel corpora that has shown to be highly productive has been technical documentation. In this way, some parallel corpora have been created, for example, from software documentation: KDE4, PHP, and OpenOffice.org (Tiedemann, 2009) are only a few examples of corpora created from this kind of documentation. Additionally, some other sources of parallel texts have been explored for which the structure of the documents cannot be used to straightforwardly align them, as in the case of translations of books, where the information can differ strongly between different editions of the same book. Some examples of such parallel corpora are the EUBookshop corpus (Skadiņš et al., 2014) or the Bilingual Books corpus.[4]

Despite the usefulness of the existing parallel corpora, the nature of the sources from which they are created still brings about a low coverage for languages with a small number of speakers or without institutions that recognise them as official languages. The growing amount of collaboratively-built data in the Internet has been a good source of parallel data during the last years, allowing the creation of corpora such as the OpenSubtitles corpus (Tiedemann, 2009), created from translations of subtitles, the Tatoeba corpus (Tiedemann, 2009), created from the data of a collaborative platform for translating sentences, or the transcription of the conferences in the website TED talks (Cettolo et al., 2012).[5] The fact that these platforms are open to the collaboration of the community of users makes it possible to find translations to languages that are not usually covered by companies or official institutions. However, despite the growing number of collaborative multilingual projects in the Internet, this kind of sources of parallel data is still limited.

As mentioned above, crawling parallel data from the Internet (Resnik and Smith, 2003) has been a recurrent solution to deal with the lack of parallel corpora for some language pairs. This technique consists in downloading texts in different

[4] http://www.farkastranslations.com/bilingual_books.php

[5] http://www.ted.com

languages from multilingual websites, detecting which of them are parallel documents, and, optionally, aligning them at the segment level. Using this technique, new parallel corpora have been created during the last years for languages with low resources (Varga et al., 2005; Mohler and Mihalcea, 2008; Tyers and Alperen, 2010). However, to the best of our knowledge, it has never been studied the impact of SBI created by means of parallel data crawling for the task of word-level QE.

1.3 Problems addressed

As mentioned above, the main goal of this dissertation is to develop methods to estimate the quality of a translation by using SBI both in TM-based CAT and MT. The motivation of the approaches described in this dissertation is to take advantage of SBI, such as bilingual dictionaries, phrase tables, TM, MT and bilingual concordancers, available, for example, in the Internet. The **main working hypothesis** is:

Main working hypothesis: it is possible to develop methods exclusively based on external SBI for word-level QE in TM-based CAT and MT.

This working hypothesis summarises the primary objective of this PhD thesis, and provides a guiding principle for this dissertation. In this section, three research blocks are described, each one formulating partial working hypotheses that help to confirm the main working hypothesis above: (i) developing methods based on SBI for word-level QE in TM-based CAT, (ii) developing methods based on SBI for word-level QE in MT, and (iii) developing methods to obtain new SBI for under-resourced language pairs. Figure 1.1 shows these research blocks (represented with white rectangles) as well as the chapter in which they are described and the relation between them. The figure also includes the reprinted publications included in each research block.

It is worth mentioning that the techniques for word-level QE developed as part of this PhD work are agnostic as to which SBI are used, that is, they make no assumptions regarding the amount, quality or source of the bilingual information used. This guarantees a high level of flexibility for the approaches developed and, in addition, allows making the most of the resources available for a given translation task.

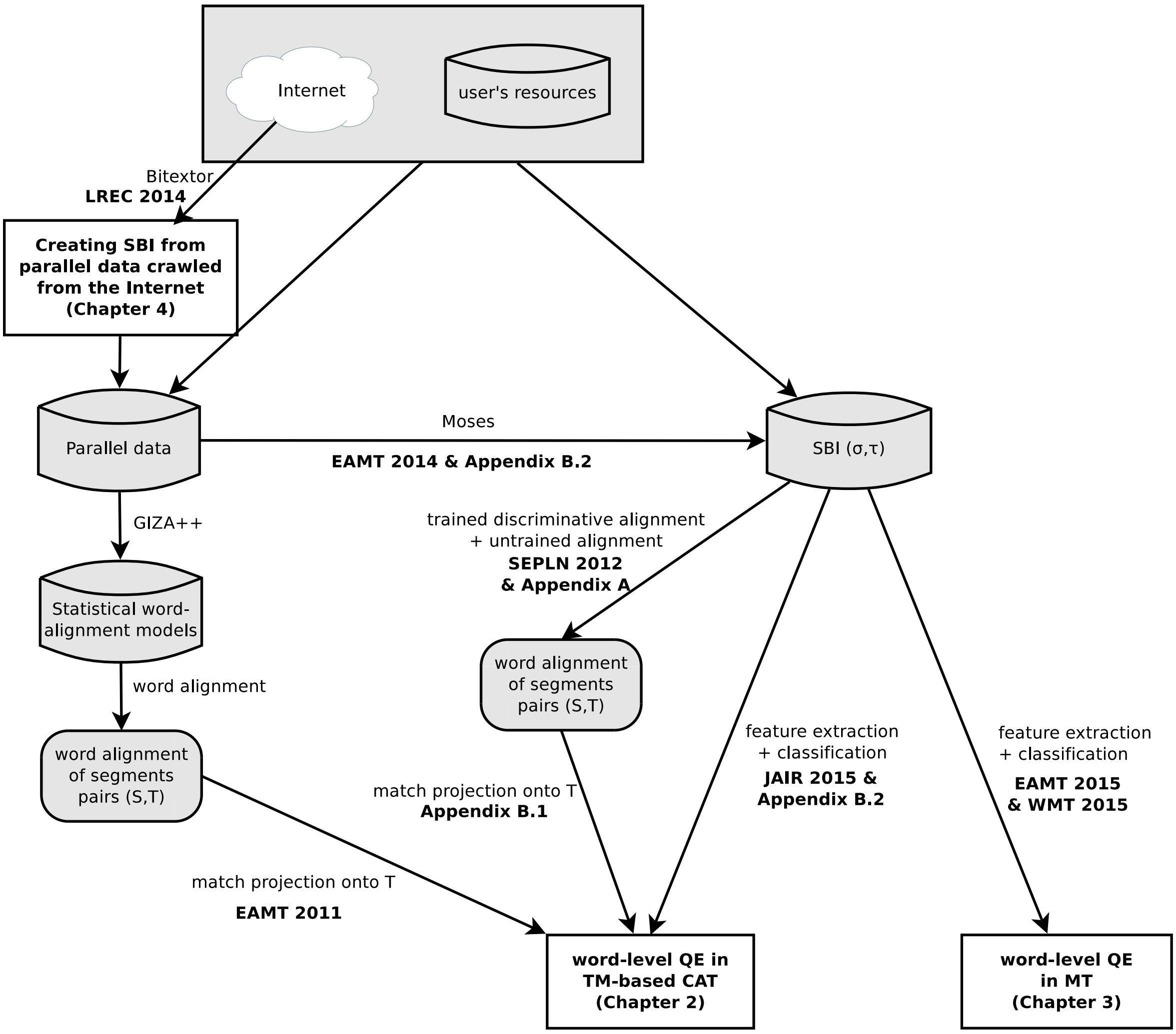

Internet
user's resources
Bitextor
LREC 2014
Creating SBI from parallel data crawled from the Internet (Chapter 4)
Parallel data
Moses
EAMT 2014 & Appendix B.2
SBI (σ,τ)
GIZA++
Statistical word-alignment models
trained discriminative alignment + untrained alignment
SEPLN 2012 & Appendix A
word alignment
word alignment of segments pairs (S,T)
word alignment of segments pairs (S,T)
feature extraction + classification
JAIR 2015 & Appendix B.2
feature extraction + classification
EAMT 2015 & WMT 2015
match projection onto T
Appendix B.1
match projection onto T
EAMT 2011
word-level QE in TM-based CAT (Chapter 2)
word-level QE in MT (Chapter 3)

Figure 1.1: This diagram represents the work developed as part of this PhD thesis, the reprinted papers included in this dissertation and the chapters in which they can be found.

1.3.1 Using sources of bilingual information for word-level quality estimation in computer-aided translation based on translation memories

The objective of this research block is to define methods for word-level QE in TM-based CAT. As it has been stated in the previous section, TM-based CAT tools usually provide information regarding the usefulness of a translation suggestion (S, T) at the word level, but only for the SL segment S. This information is obtained by comparing the segment S to the segment to be translated S' and checking which words differ between them. The main objective of providing this information to the translator is to ease the estimation of the effort required to post-edit the translation suggestion provided by the tool. However, our word-level QE approach goes a step beyond: it builds on the information about the words in S matching S' and *projects* this information onto the words in the TL segment T suggested to the translator; a task we term as *word-keeping recommendation*.

It is obvious that a method providing word-keeping recommendation directly on T would give much more information about the effort needed to complete the translation task. In addition, if this information were shown to the translator, for example, colouring the words in each translation suggestion T, it could be used as a post-editing guide: for example, words likely to be changed could be coloured in red, while words likely to be kept untouched could be coloured in green.

Despite the advantages of word-keeping recommendation, the only previous approach sharing this objective is the one described in a patent application by Kuhn et al. (2011); unfortunately, the details of this method remain unpublished. The lack of existing solutions for word-keeping recommendation may lead the reader to think that this is not an interesting task that deserves the attention of the industry or the scientific community. However, experiments performed with professional translators, which are reported in Appendix A of the reprinted publication 2.2.1 described below, confirm that high-quality word-keeping recommendation can reduce the time devoted to a translation task in up to 14%. This result highlights the relevance of the research conducted, and confirms the potential advantages for professional translators.

Chapter 2 introduces the task of word-keeping recommendation and describes a collection of methods for word-level QE in TM-based CAT. The chapter is divided in two section: Section 2.1 describes methods for word-level QE in TM-based CAT using word alignments, whereas Section 2.2 describes a method for word-level QE in TM-based CAT that uses external SBI.

The methods described in Section 2.1 use the alignment between the words in S and T to project the information about the words in S matching S' onto the words in T in order to provide word-keeping recommendations. Figure 1.2 shows an example of this technique, in which the segment in English S' to be translated is *"the Asia-Pacific Association for Machine Translation"*, and the suggested English–Catalan TU (S, T) is *("the European Association for Machine Translation", "l'Associació Europea per a la traducció automàtica")*. The figure shows how S and S' are matched to detect which words differ between both segments (in bold), and how this information is projected onto T using the word alignments represented by the edges connecting the words. This section contains two publications:

- Esplà, M., Sánchez-Martínez, F., Forcada, M.L. 2011. Using word alignments to assist computer-aided translation users by marking which target-side words to change or keep unedited. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, p. 81–89, May 30–31, 2011, Leuven, Belgium. [**Reprinted publication 2.1.1**]

- Esplà-Gomis, M., Sánchez-Martínez, F., Forcada, M.L. 2012. A simple approach to use bilingual information sources for word alignment. In *Procesamiento del Lenguaje Natural*, 49, p. 93–100. [**Reprinted publication 2.1.2**]

Reprinted publication 2.1.1 describes the techniques developed for word-level QE based on statistical word alignments (Och and Ney, 2003). The reader may have noticed that statistical word alignments cannot be considered a proper SBI according to the definition provided at the beginning of this chapter. However, it is the most obvious technique to relate the words in the segments of a TU (S, T), a step needed in order to project onto T the matching information obtained by comparing S and S'. Therefore, reprinted publication 2.1.1 is aimed at confirming the following working hypothesis:

Hypothesis #1: it is possible to use word alignments to estimate the quality of TM-based CAT suggestions at the word level.

The experiments described in reprinted publication 2.1.1, involving different translation tasks between English and Spanish, provided promising results and showed that word-keeping recommendations can be obtained with acceptable accuracy using statistical word alignments.

Reprinted publication 2.1.2 proposes a new heuristic method that determines word alignments *on the fly* by using external SBI, therefore avoiding the need to train statistical word-alignment models. This publication builds on the results of reprinted publication 2.1.1, but moving to a method based on SBI, which is one of the objectives of this work. Therefore, the working hypothesis to be confirmed is:

Hypothesis #2: it is possible to use any SBI to obtain word alignments.

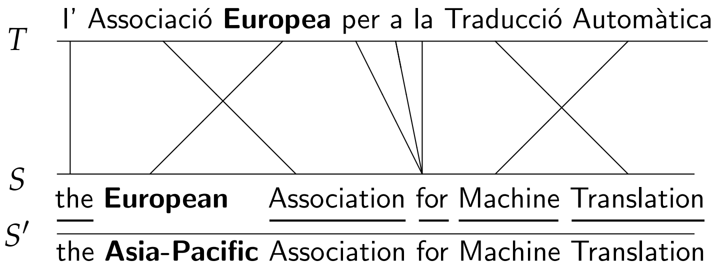

T
l' Associació **Europea** per a la Traducció Automàtica
S
the **European** Association for Machine Translation
S'
the **Asia-Pacific** Association for Machine Translation

Figure 1.2: Example of word-level QE in TM-based CAT using word alignemnts. In the example, the segment to be translated S' is compared to the SL segment S of the TU proposed by the system to detect the words in S that are not matched (in bold); word alignments are then used to extend this information to the TL segment T of the TU proposed.

Additional methods for word alignment based on SBI were explored, and they are described in Appendix A. These methods were compared to the state-of-the-art word-alignment tool GIZA++ (Och and Ney, 2003), which is based on statistical word-alignment models (Brown et al., 1993): the results obtained by the methods based on SBI were comparable to those obtained by GIZA++ in terms of precision when aligning segments between Spanish and English, but worse in terms of recall. However, the methods based on SBI obtained better results (also for recall) when small amounts of parallel data were available to train the statistical word-alignment models.

The word-alignment methods based on SBI developed open the door to obtaining word-level QE in TM-based CAT from SBI. Indeed, additional experiments are included in Appendix B in which statistical word alignments are replaced by SBI-based word alignments, in order to compare the performance of both approaches for word-level QE in TM-based CAT.

Section 2.2 describes the techniques for word-level QE based on any SBI, and contains a single publication:

- Esplà-Gomis, M., Sánchez-Martínez, F., Forcada, M.L. 2015. Using machine translation to provide target-language edit hints in computer-aided translation based on translation memories. In *Journal of Artificial Intelligence Research*, volume 53, p. 169–222. [**Reprinted publication 2.2.1**]

Reprinted publication 2.2.1 describes two different methods for word-level QE directly using SBI: an heuristic approach, and a binary classification approach based on machine learning technologies . The main objective of the research conducted in this publication is to confirm the following working hypothesis:

Hypothesis #3: it is possible to use SBI to estimate the quality of TM-based CAT translation suggestions at the word level.

The approaches described in reprinted publication 2.2.1 are compared to the previous methods based on statistical word alignment for five different language pairs: English–Spanish, English–French, English–German, English–Finnish, and Spanish–French. This offers a more reliable evaluation of the methods described in Chapter 2 because they are used to translate between closely related languages (Spanish and French are both Romance languages, while English and German are both Germanic languages), between languages in the same family (English, Spanish, German and French are all Indo-European languages), and even between languages which are not related at all (Finnish is an Uralic language). The experiments described in this article confirm that the results obtained with SBI are comparable to those obtained with statistical word alignment, and even better in some cases, in particular when the documents to be translated fall out of the domain of those used for training.

1.3.2 Using sources of bilingual information for word-level quality estimation in machine translation

The second translation technology on which this dissertation is focused is MT. Several approaches already exist for word-level QE in MT, most of them based on machine learning techniques. As mentioned in Section 1.1.1, several families of features exist that are obtained from different sources of evidence and that can be used to detect which words in a translation hypothesis need to be post-edited. However, to the best of our knowledge, none of these features use external SBI in an agnostic way. Therefore, the objective of this research block is to develop methods that, building on those described in Chapter 2, estimate the quality of translations produced by MT systems using any SBI available.

It is worth noting that there are substantial differences between the problem of word-level QE in TM-based CAT and word-level QE in MT: whereas in TM-based CAT, the problem is to detect which words in an adequate translation of S are not part of the translation of a new source segment S' to be translated, in MT one works on an automatic translation of S', which may or may not be adequate. Therefore, a substantially different approach for word-level QE in MT has been developed, but grounded on the same basic ideas as for the approach developed for TM-based CAT.

Chapter 3 describes a new approach based on SBI which tackles word-level QE in MT as a binary classification problem. This approach applies the same machine-learning methodology described in Section 2.2, but using new features to classify words in a given translation hypothesis T as "good" (to be kept) or "bad" (to be either deleted or replaced). In MT QE, two different collections of features are described: one with positive features, that is, features confirming that a given word should be part of the final translation, and one with negative features, that is, features providing evidence that a given word should be deleted or replaced. This chapter contains two publications:

- Esplà-Gomis, M., Sánchez-Martínez, F., and Forcada, M.L. 2015. Using on-line available sources of bilingual information for word-level machine translation quality estimation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, p. 19–26, Antalya, Turkey, May 11–13, 2015. [**Reprinted publication 3.1**]

- Esplà-Gomis, M., Sánchez-Martínez, F., and Forcada, M.L. 2015. UAlacant word-level machine translation quality estimation system at WMT 2015. In *Proceedings of the 10th Workshop on Statistical Machine Translation*, p. 309–315, Lisbon, Portugal, September 17–18, 2015. [**Reprinted publication 3.2**]

Both publications are aimed at confirming the following working hypothesis:

Hypothesis #4: it is possible to take the SBI-based methods for word-level QE in TM-based CAT and adapt them for their use in MT.

Reprinted publication 3.1 describes the method proposed and the collection of features used by the binary classifier. In addition, it contains a collection of experiments that evaluates the method proposed on the data provided by the organisers of the shared task on word-level MT QE in the 2014 edition of the Workshop on Statistical Machine Translation (Bojar et al., 2014).[6] The data for evaluation was provided for two language pairs, English–Spanish and German–English, in both translation directions. As will be explained in Chapter 3, despite the differences between languages, the results obtained not only confirmed the feasibility of the method, but also showed an excellent performance, comparable to that of some of the best performing approaches in the task.

Reprinted publication 3.2 in Chapter 3 describes the application of the approach proposed in this dissertation for the shared task on word-level MT QE in the 2015 edition of the Workshop on Statistical Machine Translation (Bojar et al., 2015). For this edition, evaluation data was published only for translating Spanish into English

[6] The shared task on word-level MT QE of the Workshop on Statistical Machine Translation has been organised during the last three years, and is a contest that evaluates approaches to word-level MT QE on a common data set. The organisers of the task provide a collection of source segments and their translations produced by an MT system; three data sets are provided: a training set, a development set, and a test set. For the two first sets the words in the translations are labelled as ″good″ or ″bad″ (finer labels are sometimes given to differentiate between error types), while the participants have to use their systems to label the translations in the test set. The use of a common data set provides an adequate framework to compare and evaluate the different approaches participating in the task, such as those proposed in this dissertation.

and our method, combining the features described in reprinted publication 3.1 article with the baseline features provided by the organisers, performed extremely well, ranking first (Bojar et al., 2015) among all the approaches participating in the task.

1.3.3 Building new sources of bilingual information for under-resourced language pairs by crawling parallel data

One of the main pillars of this dissertation is the availability of SBI. In fact, as explained at the beginning of this section, one of the objectives of the research conducted is to take advantage of the vast amount of SBI that are available and ready to use. However, as the reader may have guessed, this assumption is not valid for all language pairs. In fact, in Section 1.2 we already discussed about the problem of low technological coverage for many languages. The lack of SBI for under-resourced language pairs motivates this third block of research.

To deal with the shortage of SBI for some language pairs, a specific methodology was developed to harvest parallel data from multilingual websites by means of the tool Bitextor (Esplà-Gomis and Forcada, 2010) version 4.1. This tool performs document alignment based on: (i) the use of bilingual lexicons to compare the content of the files through a method based on that by Sánchez-Martínez and Carrasco (2011), and (ii) the comparison of the HTML structure of the documents (Resnik and Smith, 2003). In addition, Bitextor is able to align the documents at the segment level by using the tool Hunalign (Varga et al., 2005). Sentence-aligned data can then be used to produce SBI, such as bilingual lexicons, phrase tables or statistical MT (SMT) systems, ready to be used by the word-level QE methods described in chapters 2 and 3.

Chapter 4 describes the research conducted regarding the creation of new SBI, and contains two publications:

- Esplà-Gomis, M., Klubička, F., Ljubešić, N., Ortiz-Rojas, S., Papavassiliou, S., and Prokopidis, P. 2014. Comparing two acquisition systems for automatically building an English–Croatian parallel corpus from multilingual websites. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, p. 1252–1258, Reykjavík, Iceland, May 26–31, 2014. **[Reprinted publication 4.1]**

- Toral, A., Rubino, R., Esplà-Gomis, M., Pirinen, T., Way, A., and Ramírez-Sánchez, G. 2014. Extrinsic evaluation of web-crawlers in machine translation: a case study on Croatian–English for the tourism domain. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, p. 221–224, Dubrovnik, Croatia, June 16–18, 2014. **[Reprinted publication 4.2]**

As already mentioned in the Preface, most of the research conducted in this chapter was developed in the framework of the Abu-MaTran EU-funded project,[7] which focuses on South-Slavic languages, and pays special attention to Croatian. For this reason, both papers deal with the creation of SBI for the English–Croatian language pair. The work developed in both papers is aimed at confirming the following working hypothesis:

Hypothesis #5: it is possible to create new SBI that enable word-level QE for language pairs with no SBI available using Bitextor to crawl parallel data.

Reprinted publication 4.1 describes the intrinsic evaluation of the English–Croatian parallel data crawled from a collection of 21 websites by using both Bitextor and another state-of-the-art tool: ILSP Focused Crawler (Papavassiliou et al., 2013). The paper reports the results obtained by both tools in terms of the amount of data crawled and of the quality of the corpus built. To obtain information about the quality of the corpus crawled, a human evaluation was conducted on a fraction of it. The promising results obtained, specially regarding the quality of the parallel corpus, motivated the reprinted publication 4.2, in which an extrinsic evaluation of these data is described. This extrinsic evaluation was performed by training a phrase-based SMT system (Koehn et al., 2003) on it, and then evaluating the resulting MT system on a translation task. The results of this second paper confirmed the usefulness of the data crawled to create a fully functional SMT system.

An evaluation of the impact of these new SBI created by using Bitextor in word-level QE has not yet been published. For this reason, additional experiments were performed and are described in Appendix B. The objective of these experiments is to confirm the last working hypothesis:

Hypothesis #6: the results obtained in word-level QE for under-resourced language pairs can be improved by using new SBI obtained through parallel data crawling.

The new experiments carried out revisit previous experiments described in Section 2.2 and focus on the most under-resourced language pair included in the experiments described in reprinted publication 2.2.1: English–Finnish. The experiments in reprinted publication 2.2.1 showed that, as a result of the reduced coverage of SBI for this language pair, the quality of a large percentage of words in the test set (higher than 10%) could not be estimated. This was due to the fact that there was no information available to make it possible to project the information of the words in S matching S' onto T. The experiments described in Section B.2 prove that a dramatic improvement in the amount of words whose quality can be estimated is obtained by using new SBI created with Bitextor.

[7] `http://www.abumatran.eu`

Chapter 2

Quality estimation in computer-aided translation based on translation memories

This chapter describes the research conducted on word-level QE in TM-based CAT. The concept of word-level QE is borrowed from the field of MT, where it consists in detecting which words in the output of an MT system are adequate and which of them are inadequate. In TM-based CAT, the task consists in, given a TU (S, T) and a segment S' to be translated, finding the words in T that do not correspond to the translation of the words in S'. As this task differs from word-level QE in MT because, ideally, T is an error-free, adequate translation of S, we term this task as *word-keeping recommendation*.

This chapter includes two sections: Section 2.1 describes the heuristic methods proposed for word-keeping recommendation using word alignments; Section 2.2 describes new methods able to use any external SBI for word-keeping recommendation, therefore avoiding the need for word alignments.

2.1 Methods for word-level quality estimation in computer-aided translation based on translation memories using word alignments

As mentioned in the introduction, word alignments are one of the most straightforward ways to project information from the words in S matching S' onto T. The work described in this section is aimed at confirming the following two working hypotheses:

Hypothesis #1: it is possible to use word alignments to estimate the quality of TM-based CAT suggestions at the word level.

Hypothesis #2: it is possible to use any SBI to obtain word alignments.

Confirming working hypotheses 1 and 2 would allow us to accomplish part of the objectives of this PhD work: developing a method that uses SBI to estimate the quality in TM-based CAT at the word level. The research being reported in this section is highlighted and put in context in Figure 2.1, and is divided into two publications, each one aimed at confirming one of the two hypotheses above:

- Esplà, M., Sánchez-Martínez, F., Forcada, M.L. 2011. Using word alignments to assist computer-aided translation users by marking which target-side words to change or keep unedited. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, p. 81–89, May 30–31, 2011, Leuven, Belgium. [**Reprinted publication 2.1.1**]

- Esplà-Gomis, M., Sánchez-Martínez, F., Forcada, M.L. 2012. A simple approach to use bilingual information sources for word alignment. In *Procesamiento del Lenguaje Natural*, 49, p. 93–100. [**Reprinted publication 2.1.2**]

Reprinted publication 2.1.1 describes the methods developed for QE based on statistical word alignment. The experiments described in this article for a set of translation tasks between English and Spanish show the feasibility of the approach, obtaining an accuracy in word-keeping recommendation higher than 94% in all the translation tasks. The promising results obtained in this first work motivated the research in reprinted publication 2.1.2, which describes a new heuristic method for obtaining word alignments from external SBI, avoiding the dependency on statistical word-alignment models. This heuristic method for word alignment has an important advantage compared to other approaches to word alignment: it can be used *on the fly* to align any new pair of texts. The heuristic word-alignment method may be extended to obtain a maximum-likelihood approach, which is described in reprinted publication A.1 included in Appendix A. This maximum-likelihood aligner is more general and performs better, but it needs to be trained in advance.

Word aligners using SBI perform reasonably well when compared to state-of-the-art statistical word-alignment models such as those implemented in GIZA++ (Och and Ney, 2003). The experiments reported in this section show that, on the one hand, the precision obtained by both approaches is comparable, and on the other hand, that the recall of the word-alignment methods based on SBI is lower when GIZA++ is trained on a large parallel corpus. Despite this shortcoming, the word-alignment methods based on SBI proved to be much more convenient when no large

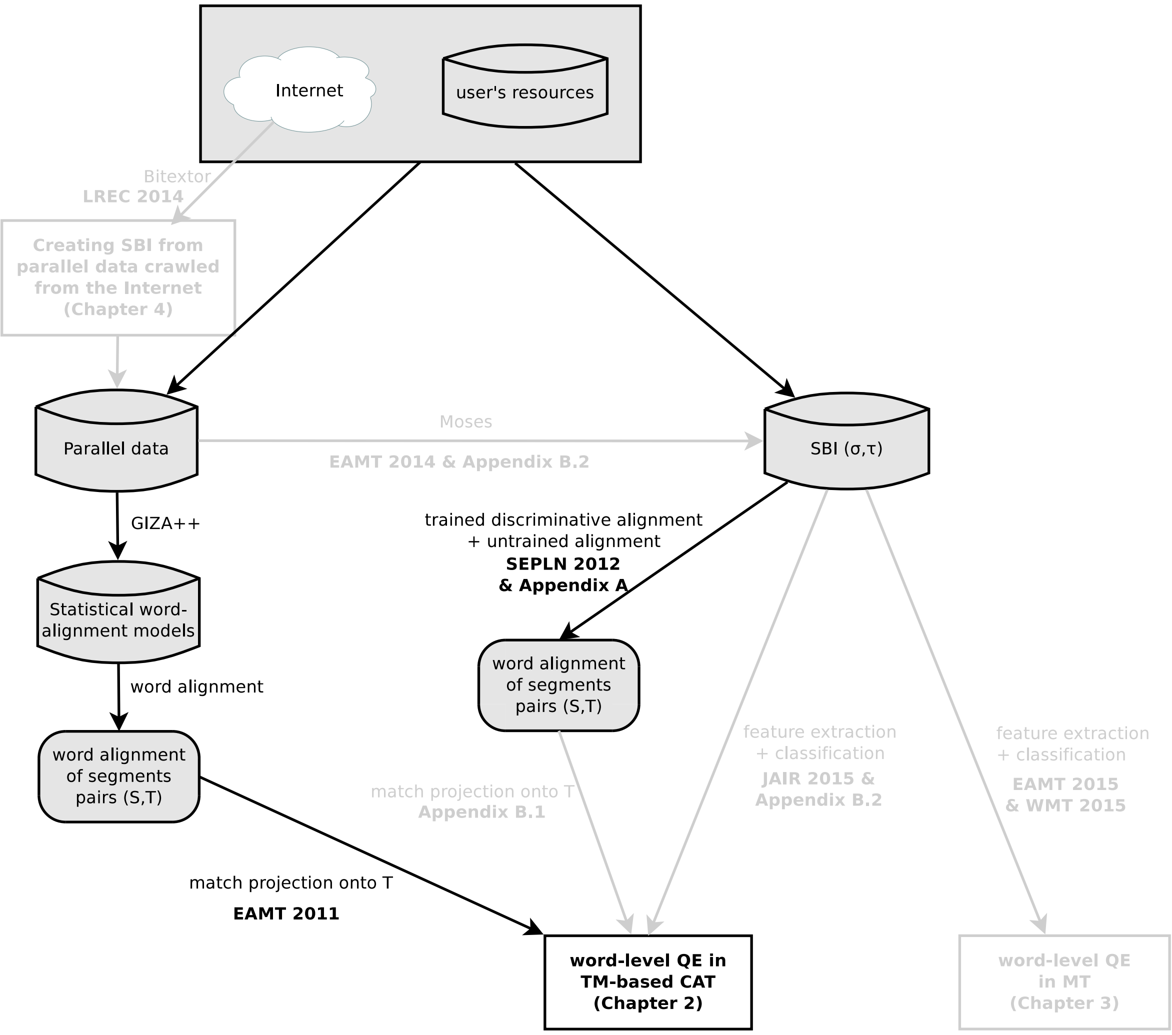

Internet
user's resources
Bitextor
LREC 2014
Creating SBI from parallel data crawled from the Internet (Chapter 4)
Parallel data
Moses
EAMT 2014 & Appendix B.2
SBI (σ,τ)
GIZA++
Statistical word-alignment models
word alignment
word alignment of segments pairs (S,T)
trained discriminative alignment + untrained alignment
SEPLN 2012 & Appendix A
word alignment of segments pairs (S,T)
match projection onto T
Appendix B.1
feature extraction + classification
JAIR 2015 & Appendix B.2
feature extraction + classification
EAMT 2015 & WMT 2015
match projection onto T
EAMT 2011
word-level QE in TM-based CAT (Chapter 2)
word-level QE in MT (Chapter 3)

Figure 2.1: This diagram highlights the research being reported in Section 2.1 (and the publications concerned) on the use of statistical word-alignment models for word-level QE in TM-based CAT, and on the use of SBI to obtain word alignments. It also shows their relation with the rest of the work reported in this dissertation.

parallel corpora are available to train statistical word-alignment models. An extrinsic evaluation of these word aligners for the task of word-level QE in TM-based CAT is provided in Appendix B.

Using word alignments to assist computer-aided translation users by marking which target-side words to change or keep unedited

Miquel Esplà and **Felipe Sánchez-Martínez** and **Mikel L. Forcada**
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain
{mespla,fsanchez,mlf}@dlsi.ua.es

Abstract

This paper explores a new method to improve computer-aided translation (CAT) systems based on translation memory (TM) by using pre-computed word alignments between the source and target segments in the translation units (TUs) of the user's TM. When a new segment is to be translated by the CAT user, our approach uses the word alignments in the matching TUs to mark the words that should be changed or kept unedited to transform the proposed translation into an adequate translation. In this paper, we evaluate different sets of alignments obtained by using GIZA++. Experiments conducted in the translation of Spanish texts into English show that this approach is able to predict which target words have to be changed or kept unedited with an accuracy above 94% for fuzzy-match scores greater or equal to 60%. In an appendix we evaluate our approach when new TUs (not seen during the computation of the word-alignment models) are used.

1 Introduction

Computer-aided translation (CAT) systems based on translation memory (TM) (Bowker, 2002; Somers, 2003) and, optionally, additional tools such as terminology databases (Bowker, 2003), are the translation technology of choice for most professional translators, especially when translation tasks are very repetitive and effective recycling of previous translations is feasible.

When using a TM-based CAT system to translate a source segment s', the system provides the set of translation units (TUs) $\{(s_i, t_i)\}_{i=1}^{N}$ whose fuzzy-match score is above a given threshold Θ, and marks which words in each source-language (SL) segment s_i differ from those in s'. It is however up to the translator to identify which target words in the corresponding target-language (TL) segments t_i should be changed to convert t_i into t', an adequate translation of s'.

The method we propose and evaluate in this paper is aimed at recommending the CAT user which words of t_i should be changed by the translator or kept unedited to transform t_i into t'. To do so, we pre-process the user's TM to compute the word alignments between the source and target segments in each TU. Then, when a new segment s' is to be translated, the TUs with a fuzzy-match score above the threshold Θ are obtained and the alignment between the words in s_i and t_i are used to mark which words in t_i should be changed or kept unedited.

Related work. In the literature one can find different approaches that use word or phrase alignments to improve existing TM-based CAT systems; although, to our knowledge, none of them use word alignments for the purpose we study in this paper. Simard (2003) focuses on the creation of TM-based CAT systems able to work at the sub-segment level by proposing as translation sub-segments extracted from longer segments in the matching TUs. To do this, he implements the *translation spotting* (Véronis and Langlais, 2000) technique by using statistical word-alignment methods (Och and Ney, 2003); translation spotting consists of identifying, for a pair of parallel sentences, the words or phrases in a TL segment that correspond to the words in a SL segment. The work by Bourdaillet et al. (2009) follows a similar approach, although it does not focus on traditional TM-based CAT systems, but

Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste (eds.)
Proceedings of the 15th Conference of the European Association for Machine Translation, p. 81–88
Leuven, Belgium, May 2011

on the use of a bilingual concordancer to assist professional translators.

More similar to our approach is the one by Kranias and Samiotou (2004) which is implemented on the *ESTeam* CAT system. Kranias and Samiotou (2004) align the source and target segments in each TU at different sub-segment levels by using a bilingual dictionary (Meyers et al., 1998), and then use these alignments to (i) identify the sub-segments in a translation proposal t_i that need to be changed, and (ii) propose a machine translation for them.

In this paper we propose a different way of using word alignments in a TM-based CAT system to alleviate the task of professional translators. The main difference between our approach and those previously described is that in our approach word alignments are used only to recommend the words to be changed or kept unedited, without proposing a translation for them, so that the user can focus on choosing a translation where words have to be changed. It is worth noting that as we do not change the translation proposals in any way, our approach does not affect the predictability of TM proposals and the way in which fuzzy-match scores (Sikes, 2007) are interpreted by the CAT user. In addition, our system is independent of any external resources, such as MT systems or dictionaries, as opposed to the work by Kranias and Samiotou (2004).

The rest of the paper is organized as follows. Section 2 presents the way in which word alignments are used by our approach and the different word alignment methods we have tried. Section 3 then describes the experimental framework, whereas Section 4 discusses the results obtained. Section 5 includes some concluding remarks and plans for future work. In Appendix A we evaluate our approach when it is applied to new TUs not seen during the computation of the word-alignment models used.

2 Methodology

Let w_{ij} be the word in the j-th position of segment t_i which is aligned with word v_{ik} in the k-th position of its counterpart segment s_i. If v_{ik} is part of the match between s_i and s' (the new segment to be translated), then this indicates that w_{ij} might be part of the translation of that word and, therefore, it should be kept unedited. Conversely, if $v_{ik'}$ is not part of the match between s_i and s', this indicates that $w_{ij'}$ might not be the translation of any of the words in s' and it should be changed (see Figure 1). Note that w_{ij} may not be aligned with any word in s_i, and that in these cases nothing can be said about it. This information may be shown using colour codes, for example, red for the words to be changed, green for the words to be kept unedited and yellow for those unaligned words for which nothing can be said.

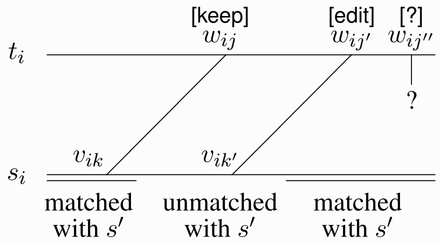


Figure 1: Target word w_{ij} may have to be kept unedited because it is aligned with source word v_{ik} which is in the part of s_i that matches s'. Target word $w_{ij'}$ may have to be changed because it is aligned with source word $v_{ik'}$ which is in the part of s_i that does not match s'. As target word $w_{ij''}$ is not aligned to any source word in s_i, nothing can be said about it.

To determine if word w_{ij} in the target proposal t_i should be changed or kept unedited, we compute the fraction of words v_{ik} aligned to w_{ij} which are common to both s_i and s':

$$f_K(w_{ij}, s', s_i, t_i) = \frac{\sum_{v_{ik} \in \text{aligned}(w_{ij})} \text{matched}(v_{ik})}{|\text{aligned}(w_{ij})|}$$

where $\text{aligned}(w_{ij})$ is the set of source words in s_i which are aligned with target word w_{ij} in t_i, and $\text{matched}(v_{ik})$ equals 1 if word v_{ik} is part of the match between s_i and s', the segment to be translated, and 0 otherwise. Function $\text{matched}(x)$ is based on the optimal edit path, obtained as a result of the word-based Levenshtein distance (Levenshtein, 1966) between the segment to be translated and the SL segment of the matching TU. The fraction $f_K(w_{ij}, s', s_i, t_i)$ may be interpreted as the likelihood that word w_{ij} has to be kept unedited. If $|\text{aligned}(w_{ij})|$ happens to be zero, $f_K(w_{ij}, s', s_i, t_i)$ is arbitrarily set to $\frac{1}{2}$, meaning "do not know".

We have chosen the likelihood that word w_{ij} will be kept unedited to depend on how many SL words aligned with it are matched with the SL segment to be translated. It may happen that w_{ij} is aligned with one or more words in s_i that are matched with words in s', and, at the same time, it is aligned with one or more unmatched words in s_i. In the experiments we have tried two ways of dealing with this, one that requires all SL words in s_i to be matched, and another one that only requires the majority of words aligned with w_{ij} to be matched. These strategies have been chosen because of their simplicity, although it could also be possible to use, for example, a maximum entropy classifier (Berger et al., 1996), in order to determine which words should be changed or kept unedited. In that case, f_K would be one of the features used by the maximum entropy classifier.

To illustrate these ideas, Figure 2 shows an example of a word-aligned pair of segments (s_i and t_i) and a segment s' to be translated. As can be seen, the word *he* in t_i is aligned with the word *él* in s_i, which does not match with any word in s'. Therefore, *he* should be marked to be changed. Conversely, the words *his* and *brother* are aligned with *su* and *hermano*, respectively, which are matched in s' and, therefore should be kept unedited. Finally, the word *missed* is aligned with three words in s_i: *echó* and *de*, which are matched in s', and *menos*, which is not matched. In this case, if the criterion of unanimity is applied, the word would be marked neither as "keep" nor as "change". Otherwise, if the criterion of majority is applied, the word would be marked to be changed.

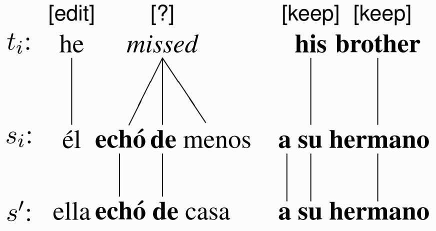

[edit]
[?]
[keep]
[keep]
t_i: he *missed* **his brother**
s_i: él **echó de** menos **a su hermano**
s': ella **echó de** casa **a su hermano**

Figure 2: Example of alignment and matching.

For the experiments in this paper we have used word alignments obtained by means of the free/open-source GIZA++[1] tool (Och and Ney, 2003) which implements standard word-based statistical machine translation models (Brown et al., 1993) as well as a hidden-Markov-model-based alignment model (Vogel et al., 1996). GIZA++ produces alignments in which a source word can be aligned with many target words, whereas a target word is aligned with, at most, one source word. Following common practice in statistical machine translation (Koehn, 2010, Ch. 4) we have obtained a set of symmetric word alignments by running GIZA++ in both translation directions, and then symmetrizing both sets of alignments. In the experiments we have tried the following symmetrization methods:

- the union of both sets of alignments,
- the intersection of the two alignment sets, and
- the use of the *grow-diag-final-and* heuristic (Koehn et al., 2003) as implemented in Moses (Koehn et al., 2007).

3 Experimental settings

We have tested our approach in the translation of Spanish texts into English by using two TMs: $\mathrm{TM}_{\mathrm{trans}}$ and $\mathrm{TM}_{\mathrm{test}}$. Evaluation was carried out by simulating the translation of the SL segments in $\mathrm{TM}_{\mathrm{trans}}$ by using the TUs in $\mathrm{TM}_{\mathrm{test}}$. We firstly obtained the word alignments between the parallel segments of $\mathrm{TM}_{\mathrm{test}}$ by training and running GIZA++ on the TM itself. Then, for each source segment in $\mathrm{TM}_{\mathrm{trans}}$, we obtained the TUs in $\mathrm{TM}_{\mathrm{test}}$ having a fuzzy-match score above threshold Θ, and tagged the words in their target segments as "keep" or "change".

3.1 Fuzzy-match score function

As in most TM-based CAT systems, we have chosen a fuzzy-match score function based on the Levenshtein distance (Levenshtein, 1966):

$$\mathrm{score}(s', s_i) = 1 - \frac{\mathrm{D}(s', s_i)}{\max(|s'|, |s_i|)}$$

where $|x|$ stands for the length (in words) of string x and $\mathrm{D}(x, y)$ refers to the word-based Levenshtein distance (edit distance) between x and y.

3.2 Corpora

The TMs we have used were extracted from the JRC-Acquis corpus version 3 (Steinberger et al., 2006),[2] which contains the total body of European Union (EU) law. Before extracting the TMs used, this corpus was tokenized and lowercased, and then segment pairs in which either of the segments was empty or had more than 9 times words than its counterpart were removed. Finally, segments longer than 40 words (and their corresponding counterparts) were removed because of the inability of GIZA++ to align longer segments.

[1] `http://code.google.com/p/giza-pp/`

[2] `http://wt.jrc.it/lt/Acquis/`

$\Theta(\%)$	TUs	N_{words}
50	9.5	484,523
60	6.0	303,193
70	4.5	220,304
80	3.5	166,762
90	0.9	42,708

Table 1: Average number of matching TUs per segment and number of words to tag for different fuzzy-match score thresholds (Θ).

Finally, the segment pairs in TM_{trans} and TM_{test} were randomly chosen without repetition from the resulting corpus. TM_{test} consists of 10,000 parallel segments, whereas TM_{trans} consists of 5,000 segment pairs. It is worth noting that these TMs may contain incorrect TUs as a result of wrong segment alignments and this can negatively affect the results obtained.

With respect to the number of TUs found in TM_{test} when simulating the translation of the SL segments in TM_{trans}, Table 1 reports for the different fuzzy-match score thresholds we have used: the averaged number of TUs per segment to be translated and the total number of words to classify as "keep" or "change". These data provide an idea of how repetitive the corpora we have used to carry out the experiments are.

3.3 Evaluation

We evaluate our approach for different fuzzy-match score thresholds Θ by computing the accuracy, i.e. the percentage of times the recommendation of our system is correct, and the coverage, i.e. the percentage of words for which our system is able to say something. For that purpose we calculate the optimal edit path between the target segments in TM_{trans} and the translation proposals in TM_{test} to determine the actual word-editing needs in each translation proposal.

For each SL segment s' in TM_{trans} we compute the set of matching TUs $\{(s_i, t_i)\}_{i=1}^{N}$ in TM_{test} whose fuzzy-match score is above threshold Θ. We then calculate the fraction $f_K(w_{ij}, s', s_i, t_i)$ representing the likelihood that word w_{ij} in t_i will be kept unedited and use it to mark w_{ij} as having to be changed or kept unedited by using the two different criteria (unanimity or majority) mentioned above:

unanimity: if $f_K(\cdot) = 1$ the word is tagged as "keep", whereas if $f_K(\cdot) = 0$ it is tagged as "change"; in the rest of cases no recommendation is made for that word.

majority: if $f_K(\cdot) > 0.5$ the word is tagged as "keep", whereas it is tagged as "change" if $f_K(\cdot) < 0.5$; in the unlikely case of having $f_K(\cdot) = 0.5$ no recommendation is made about that word.

The first criterion requires all the source words aligned with word w_{ij} to be matched (conversely, unmatched) with a word in the new segment to be translated, while the second criterion only requires the majority or source words aligned with w_{ij} to be matched (conversely, unmatched).

4 Results and discussion

We evaluated our approach with the different sets of word alignments obtained through the symmetrization methods described in Section 2 for values of the fuzzy-match score threshold Θ between 50% and 90%.

Tables 2 and 3 reports the accuracy and the coverage obtained with each set of alignments together with their confidence intervals for a statistical significance level $p = 0.99$ (DeGroot and Schervish, 2002, Sec. 7.5) when the majority criterion and the unanimity criterion, respectively, are used to mark the words as "keep" or "change".

As can be seen, with both criteria the best accuracy is achieved with the set of alignments obtained through the intersection method, although the use of this set of alignments shows a smaller coverage as compared to the other two sets of alignments. The use of either the union or the grow-diag-final-and sets of alignments seems to have a small impact on the accuracy although the coverage obtained for the union is slightly better. Note that with the alignments obtained by means of the intersection method, both criteria are equivalent because each word is aligned at most with one word in the other language.

The use of the unanimity criterion causes the accuracy to grow slightly, as compared to the majority criterion, while the coverage gets slightly worse as expected. It is worth noting that for fuzzy-match score thresholds above 50% differences in accuracy between both criteria are insignificant, whereas the differences in coverage are small, but significant for values of 60% and 70% of Θ.

Finally, it is important to remark that for values of Θ greater or equal to 60%, which are the values that professional translators tend to use (Bowker,

Θ (%)	Union		Intersection		GDFA	
	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)
50	92.35 ± .10	97.33 ± .06	93.80 ± .10	90.78 ± .11	92.34 ± .10	96.73 ± .07
60	94.62 ± .11	98.06 ± .07	95.80 ± .10	92.44 ± .12	94.72 ± .11	97.70 ± .07
70	97.19 ± .10	98.69 ± .06	98.04 ± .08	94.03 ± .13	97.31 ± .09	98.37 ± .07
80	98.31 ± .08	99.05 ± .06	98.82 ± .07	95.50 ± .13	98.44 ± .08	98.78 ± .07
90	97.97 ± .18	99.24 ± .11	98.75 ± .14	95.41 ± .26	98.25 ± .16	98.75 ± .14

Table 2: For different fuzzy-match score thresholds (Θ), accuracy (Acc.) and coverage (Cover.) obtained by the majority criterion for the three different sets of word alignments: intersection, union and grow-diag-final-and (GDFA).

Θ (%)	Union		Intersection		GDFA	
	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)
50	92.53 ± .10	96.87 ± .06	93.80 ± .10	90.78 ± .11	92.43 ± .10	96.50 ± .07
60	94.73 ± .11	97.78 ± .07	95.80 ± .10	92.44 ± .12	94.76 ± .11	97.57 ± .07
70	97.26 ± .10	98.50 ± .07	98.04 ± .08	94.03 ± .13	97.35 ± .09	98.30 ± .07
80	98.35 ± .08	98.96 ± .06	98.82 ± .07	95.50 ± .13	98.45 ± .08	98.75 ± .07
90	98.02 ± .18	99.17 ± .11	98.75 ± .14	95.41 ± .26	98.26 ± .16	98.73 ± .14

Table 3: For different fuzzy-match score thresholds (Θ), accuracy (Acc.) and coverage (Cover.) obtained by the unanimity criterion for the three different sets of word alignments: intersection, union and grow-diag-final-and (GDFA).

2002, p. 100), with the three sets of alignments and with both criteria accuracy is always above 94%.

5 Concluding remarks

In this paper we have presented and evaluated a new approach to guide TM-based CAT users by recommending the words in a translation proposal that should be changed or kept unedited. The method we propose requires the TM to be pre-processed in advance in order to get the alignment between the words in the source and target segments of the TUs. In any case, this pre-processing needs to be done only once, although to consider new TUs created by the user it may be worth to re-run the alignment procedure (see Appendix A). The experiments conducted in the translation of Spanish texts into English show an accuracy above 94% for fuzzy-match score thresholds greater or equal to 60% and above 97% for fuzzy-match score thresholds above 60%.

Our approach is intended to guide the TM-based CAT user in a seamless way, without distorting the known advantages of the TM-based CAT systems, namely, high predictability of the translation proposals and easy interpretation of fuzzy-match scores. We plan to field-test this approach with professional translators in order to measure the possible productivity improvements. To do this we will integrate this method in OmegaT,[3] a free/open-source TM-based CAT system.

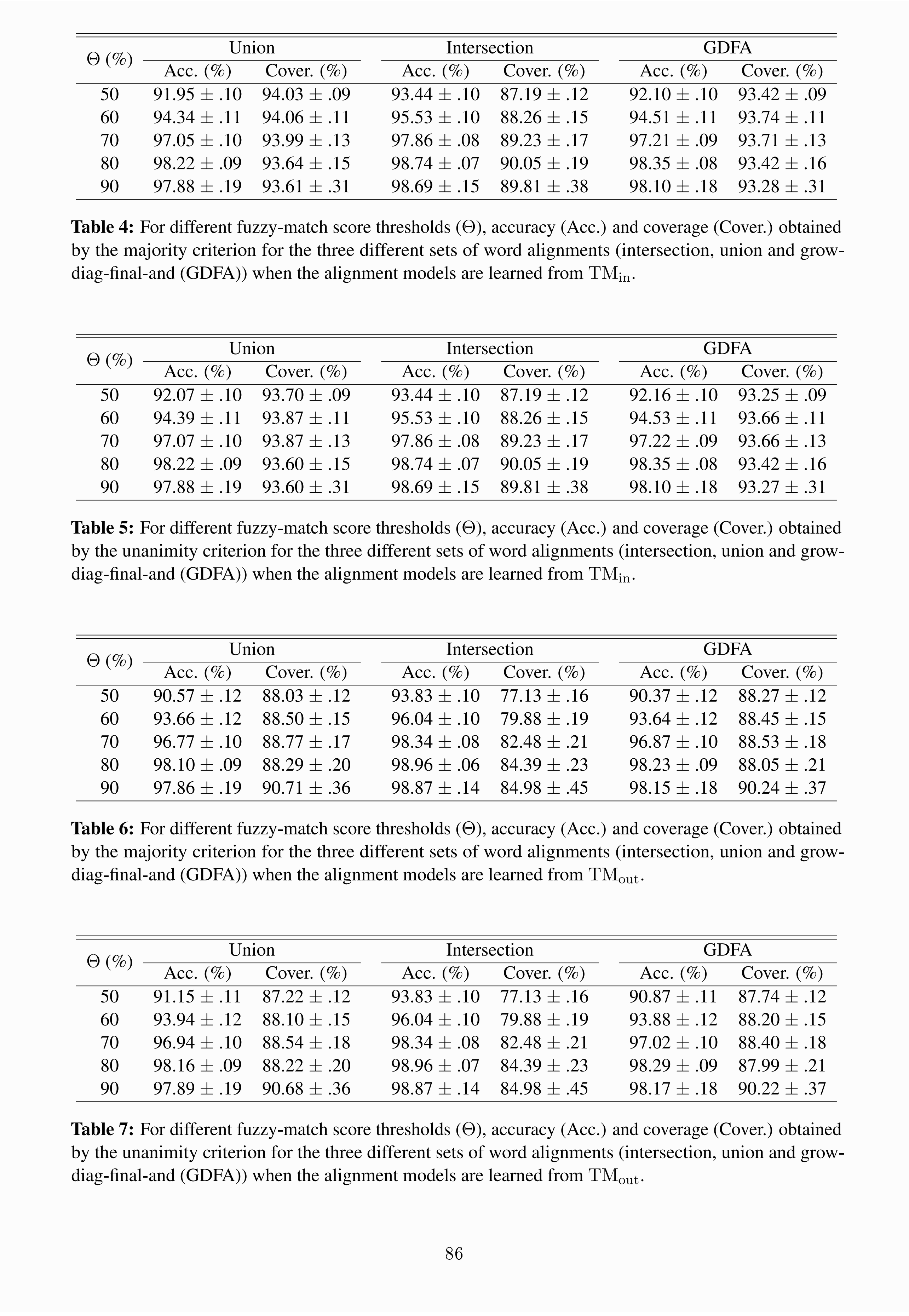
A Adding new TUs to the TM

In our experiments, we obtained the word alignment models from $\mathrm{TM}_{\mathrm{test}}$ and used them to align the words in the TUs of the same TM. In this way, we used the most information available to obtain the best word alignments possible. However, TMs are not always static and new TUs can be added to them during a translation job. In this case, the previously computed alignment models could be less effective to align the segments in the new TUs.

In this appendix, we evaluate the re-usability of previously computed alignment models on new TUs for our approach. To do so, we used an in-domain TM ($\mathrm{TM}_{\mathrm{in}}$) and an out-of-domain TM ($\mathrm{TM}_{\mathrm{out}}$) to obtain the alignment models and used them to align the segments in the TUs of $\mathrm{TM}_{\mathrm{test}}$. We then repeated the same experiments described in Section 3.3 in order to compare the results obtained.

$\mathrm{TM}_{\mathrm{in}}$ was built with 10,000 pairs of segments extracted from the JCR-Acquis corpus. These pairs of segments were chosen so as to avoid any common TU between $\mathrm{TM}_{\mathrm{in}}$ and $\mathrm{TM}_{\mathrm{test}}$, or between

[3] `http://www.omegat.org`



Θ (%)	Union		Intersection		GDFA	
	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)
50	91.95 ± .10	94.03 ± .09	93.44 ± .10	87.19 ± .12	92.10 ± .10	93.42 ± .09
60	94.34 ± .11	94.06 ± .11	95.53 ± .10	88.26 ± .15	94.51 ± .11	93.74 ± .11
70	97.05 ± .10	93.99 ± .13	97.86 ± .08	89.23 ± .17	97.21 ± .09	93.71 ± .13
80	98.22 ± .09	93.64 ± .15	98.74 ± .07	90.05 ± .19	98.35 ± .08	93.42 ± .16
90	97.88 ± .19	93.61 ± .31	98.69 ± .15	89.81 ± .38	98.10 ± .18	93.28 ± .31

Table 4: For different fuzzy-match score thresholds (Θ), accuracy (Acc.) and coverage (Cover.) obtained by the majority criterion for the three different sets of word alignments (intersection, union and grow-diag-final-and (GDFA)) when the alignment models are learned from $\mathrm{TM}_{\mathrm{in}}$.

Θ (%)	Union		Intersection		GDFA	
	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)
50	92.07 ± .10	93.70 ± .09	93.44 ± .10	87.19 ± .12	92.16 ± .10	93.25 ± .09
60	94.39 ± .11	93.87 ± .11	95.53 ± .10	88.26 ± .15	94.53 ± .11	93.66 ± .11
70	97.07 ± .10	93.87 ± .13	97.86 ± .08	89.23 ± .17	97.22 ± .09	93.66 ± .13
80	98.22 ± .09	93.60 ± .15	98.74 ± .07	90.05 ± .19	98.35 ± .08	93.42 ± .16
90	97.88 ± .19	93.60 ± .31	98.69 ± .15	89.81 ± .38	98.10 ± .18	93.27 ± .31

Table 5: For different fuzzy-match score thresholds (Θ), accuracy (Acc.) and coverage (Cover.) obtained by the unanimity criterion for the three different sets of word alignments (intersection, union and grow-diag-final-and (GDFA)) when the alignment models are learned from $\mathrm{TM}_{\mathrm{in}}$.

Θ (%)	Union		Intersection		GDFA	
	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)
50	90.57 ± .12	88.03 ± .12	93.83 ± .10	77.13 ± .16	90.37 ± .12	88.27 ± .12
60	93.66 ± .12	88.50 ± .15	96.04 ± .10	79.88 ± .19	93.64 ± .12	88.45 ± .15
70	96.77 ± .10	88.77 ± .17	98.34 ± .08	82.48 ± .21	96.87 ± .10	88.53 ± .18
80	98.10 ± .09	88.29 ± .20	98.96 ± .06	84.39 ± .23	98.23 ± .09	88.05 ± .21
90	97.86 ± .19	90.71 ± .36	98.87 ± .14	84.98 ± .45	98.15 ± .18	90.24 ± .37

Table 6: For different fuzzy-match score thresholds (Θ), accuracy (Acc.) and coverage (Cover.) obtained by the majority criterion for the three different sets of word alignments (intersection, union and grow-diag-final-and (GDFA)) when the alignment models are learned from $\mathrm{TM}_{\mathrm{out}}$.

Θ (%)	Union		Intersection		GDFA	
	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)	Acc. (%)	Cover. (%)
50	91.15 ± .11	87.22 ± .12	93.83 ± .10	77.13 ± .16	90.87 ± .11	87.74 ± .12
60	93.94 ± .12	88.10 ± .15	96.04 ± .10	79.88 ± .19	93.88 ± .12	88.20 ± .15
70	96.94 ± .10	88.54 ± .18	98.34 ± .08	82.48 ± .21	97.02 ± .10	88.40 ± .18
80	98.16 ± .09	88.22 ± .20	98.96 ± .07	84.39 ± .23	98.29 ± .09	87.99 ± .21
90	97.89 ± .19	90.68 ± .36	98.87 ± .14	84.98 ± .45	98.17 ± .18	90.22 ± .37

Table 7: For different fuzzy-match score thresholds (Θ), accuracy (Acc.) and coverage (Cover.) obtained by the unanimity criterion for the three different sets of word alignments (intersection, union and grow-diag-final-and (GDFA)) when the alignment models are learned from $\mathrm{TM}_{\mathrm{out}}$.

$\mathrm{TM_{in}}$ and $\mathrm{TM_{trans}}$. $\mathrm{TM_{out}}$ was built with 10,000 pairs of segments extracted from the EMEA corpus version 0.3 (Tiedemann, 2009),[4] which is a compilation of documents from the European Medicines Agency, and, therefore, it clearly belongs to a different domain. Before extracting the TUs, the EMEA corpus was pre-processed in the same way that the JRC-Acquis was (see Section 3.2).

Tables 4 and 5 show the results of the experiments when using the alignment models learned from $\mathrm{TM_{in}}$ for the majority criterion and for the unanimity criterion, respectively. Analogously, tables 6 and 7 show the analogous results when the alignment models learned from $\mathrm{TM_{out}}$ are used.

As can be seen, the accuracy obtained by our approach when re-using alignment models from an in-domain corpus is very similar to that obtained when these alignments are learned from the TM whose TUs are aligned. Even when the alignment models are learned from an out-of-domain corpus, the loss of accuracy is, in the worst case, lower than 2%. The main problem is the loss of coverage, which is about 6% for the in-domain training and higher than a 10% for the out-of-domain training.

On the one hand, these results show that our approach is able to re-use alignment models computed for a TM on subsequently added TUs keeping a reasonable accuracy in the recommendations. On the other hand, it is obvious that our method becomes less informative for these new TUs as their domain differs from the domain from which the alignment models have been learned.

Acknowledgements: Work supported by Spanish government through project TIN2009-14009-C02-01. The authors thank Yanjun Ma, Andy Way and Harold Somers for suggestions.

References

Berger, A.L., V.J. Della Pietra, and S.A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Bourdaillet, J., S. Huet, F. Gotti, G. Lapalme, and P. Langlais. 2009. Enhancing the bilingual concordancer TransSearch with word-level alignment. In *Proceedings of the 22nd Canadian Conference on Artificial Intelligence*, volume 5549 of *Lecture Notes in Artificial Intelligence*, pages 27–38. Springer.

Bowker, L., 2002. *Computer-aided translation technology: a practical introduction*, chapter Translation-Memory Systems, pages 92–127. University of Ottawa Press.

Bowker, L. 2003. Terminology tools for translators. In Somers, H., editor, *Computers and Translation: A Translator's Guide*, pages 49–65. John Benjamins.

Brown, P.F., S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

DeGroot, M. H. and M. J. Schervish. 2002. *Probability and Statistics*. Addison-Wesley, third edition.

Koehn, P., F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180, Prague, Czech Republic.

Koehn, P. 2010. *Statistical Machine Translation*. Cambridge University Press.

Kranias, L. and A. Samiotou. 2004. Automatic translation memory fuzzy match post-editing: A step beyond traditional TM/MT integration. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 52–57, Lisbon, Portugal.

Levenshtein, V.I. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady.*, 10(8):707–710.

Meyers, A., M. Kosaka, and R. Grishman. 1998. A multilingual procedure for dictionary-based sentence alignment. In *Machine Translation and the Information Soup*, volume 1529 of *Lecture Notes in Computer Science*, pages 187–198. Springer.

Och, F.J. and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Sikes, R. 2007. Fuzzy matching in theory and practice. *MultiLingual*, 18(6):39–43.

Simard, M. 2003. Translation spotting for translation memories. In *Proceedings of the HLT-NAACL 2003, Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 65–72, Morristown, NJ, USA.

Somers, H., 2003. *Computers and translation: a translator's guide*, chapter Translation Memory Systems, pages 31–48. John Benjamins.

[4] http://opus.lingfil.uu.se/EMEA.php

Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, and D. Tufiş. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 2142–2147, Genoa, Italy.

Tiedemann, J. 2009. News from OPUS - a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Borovets, Bulgaria.

Vogel, S., H. Ney, and C. Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 836–841, Copenhagen, Denmark.

Véronis, J. and P. Langlais, 2000. *Parallel Text Processing: Alignment and Use of Translation Corpora (Text, Speech and Language Technology)*, chapter Evaluation of Parallel Text Alignment Systems – The ARCADE Project, pages 369–388. Kluwer Academic Publishers.

A Simple Approach to Use Bilingual Information Sources for Word Alignment

Una manera sencilla para usar fuentes de información bilingüe para el alineamiento de palabras

Miquel Esplà-Gomis, Felipe Sánchez-Martínez, Mikel L. Forcada
Dep. de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain
{mespla,fsanchez,mlf}@dlsi.ua.es

Resumen: En este artículo se describe un método nuevo y sencillo para utilizar fuentes de información bilingüe para el alineamiento de palabras en segmentos de texto paralelos. Este método puede ser utilizado *al vuelo*, ya que no requiere de entrenamiento. Además, puede ser utilizado con corpus comparables. Hemos comparado los resultados de nuestro método con los obtenidos por la herramienta GIZA++, ampliamente utilizada para el alineamiento de palabras, obteniendo unos resultados bastante similares.
Palabras clave: Alineamiento de palabras, fuentes de información bilingüe

Abstract: In this paper we present a new and simple method for using sources of bilingual information for word alignment between parallel segments of text. This method can be used *on the fly*, since it does not need to be trained. In addition, it can also be applied on comparable corpora. We compare our method to the state-of-the-art tool GIZA++, widely used for word alignment, and we obtain very similar results.
Keywords: Word alignment, sources of bilingual information

1 Introduction

In this paper we describe a method which uses sources of bilingual information (SBI) such as lexicons, translation memories, or machine translation, to align the words of a segment with those in its translation (parallel segments) without any training process. Our approach aligns the sub-segments in a pair of segments S and T by using any SBI available, and then aligns the words in S and T by using a heuristic method which does not require the availability of a parallel corpus. It is worth noting that many SBIs which could be used to align words with our method are currently freely available in the Internet: MT systems, such as Apertium[1] or Google Translate;[2] bilingual dictionaries, such as Dics.info;[3] or Word Refference[4] or translation memories, such as Linguee[5] or MyMemory.[6] This method is inspired on a previous approach (Esplà-Gomis, Sánchez-Martínez, and Forcada, 2011) that was proposed to detect sub-segment alignments (SSAs) and help translators to edit the translation proposals produced by translation-memory-based computer-aided translation tools by suggesting the target words to change. A similar technique was also successfully applied to cross-lingual textual entailment detection (Esplà-Gomis, Sánchez-Martínez, and Forcada, 2012). Here, we propose to use these SSAs to obtain word alignment *on the fly*.

Related works. Many previous works tackle the problem of word alignment. The existing approaches may be divided in statistical approaches and heuristic approaches. One of the most remarkable works in the first group is the one by Brown et al. (1993), which describes a set of methods for word alignment based on the expectation-maximisation algorithm (Dempster, Laird, and Rubin, 1977), usually called *IBM models*. In this work, au-

[1] http://www.apertium.org
[2] http://translate.google.com
[3] http://www.dics.info
[4] http://www.wordreference.com
[5] http://www.linguee
[6] http://mymemory.translated.net

ISSN 1135-5948

thors propose five models, from a very simple one considering just one-to-one alignments between words, to more complex models which allow a word to be aligned with many words. Other authors (Vogel, Ney, and Tillmann, 1996; Dagan, Church, and Gale, 1993) propose using a hidden Markov model for word alignment. Both methods were combined and extended by Och and Ney (2003), who also developed the tool GIZA++, implementing all these methods.

Some heuristic approaches have also been proposed. Rapp (1999) proposes an approach based in the idea that groups of words which usually appear together in a language should also appear together in other languages. To obtain word alignments from this idea, the author uses a window of a given number of words to look for the most usual groups of words in each monolingual corpora. Then, coocurrence vectors are computed for the words appearing frequently together inside the window and word alignments are computed by comparing these coocurrence vectors. Fung and McKeown (1997) propose a similar method which introduces some SBIs. In this case, authors use bilingual dictionaries to obtain an initial alignment between *seed words* in a parallel text. To choose reliable seed words, they use only those words having a univocal translation in both directions and appearing with enough frequency to become useful references in both texts of the parallel corpus. Then, these initial alignments are used to align other words appearing around them in the parallel texts using a similar method to that used by Rapp (1999). Another family of heuristic methods for word alignment are based on cognates. Schulz et al. (2004) use word similarity between Spanish and Portuguese for word alignment. The most important limitation of this work is that it is only useful for closely-related languages. Other works (Al-Onaizan and Knight, 2002) try to overcome this problem by using transliteration to obtain the way in which a word in a language may be written in another language. In this case, Al-Onaizan and Knight (2002) use transliteration to find out the most likely way in which English proper nouns could be written in languages such as Arabic or Japanese in order to find their translations. Although statistical approaches have proved to obtain better results than heuristic ones, one of the advantages of heuristic approaches is that they can be used not only on parallel corpora, but in comparable corpora.

Novelty. In this work we propose a method for word alignment using previously existing bilingual resources. Although some works in the bibliography also use SBIs to perform alignment (Fung and McKeown, 1997), the main difference between this work and the previous approaches is that our method does not need any training process or bilingual corpus, i.e. it can be run *on the fly* on a pair of parallel segments. This kind of alignment method may be useful in some scenarios, as is the case of some computer-aided translation systems, to help users to detect which words should be post-edited in the translation proposals (Kranias and Samiotou, 2004; Esplà, Sánchez-Martínez, and Forcada, 2011). In addition, this method can be applied on comparable corpora to find partial alignments.

The paper is organized as follows: Section 2 describes the method used to collect the bilingual information and obtain the word alignment; Section 3 explains the experimental framework; Section 4 shows the results obtained for the different features combination proposed; finally, the paper ends with some concluding remarks.

2 Methodology

The method presented here uses the available sources of bilingual information (SBIs) to detect parallel sub-segments in a given pair of parallel text segments S and T written in different languages. Once sub-segments have been aligned, a simple heuristic method is used to extract the most likely word alignments from S to T and from T to S. Finally, both alignments are symmetrised to obtain the word alignments.

Sub-segment alignment. To obtain the sub-segment alignments, both segments S and T are segmented in all possible ways to obtain sub-segments of length $l \in [1, L]$, where L is a given maximum sub-segment length measured in words. Let σ be a sub-segment from S and τ a sub-segment from T. We consider that σ and τ are aligned if any of the available SBIs confirm that σ is a translation of τ, or vice versa.

Suppose the pair of parallel segments S=*Costarà temps solucionar el problema*, in Catalan, and T=*It will take time to solve the problem*, in English. We first obtain all the possible sub-segments σ in S and τ in T and then use machine translation (MT) as a SBI by translating the sub-segments in both directions. We obtain the following set of SSAs:

temps	$\leftrightarrow$	*time*
problema	$\leftrightarrow$	*problem*
solucionar el	$\rightarrow$	*solve the*
solucionar el	$\leftarrow$	*to solve the*
el problema	$\leftrightarrow$	*the problem*

It is worth noting that multiple alignments for a sub-segment are possible, as in the case of the sub-segment *solucionar el* which is both aligned with *solve the* and *to solve the.* In those cases, all the sub-segment alignments available are used. Figure 1 shows a graphical representation of these alignments.

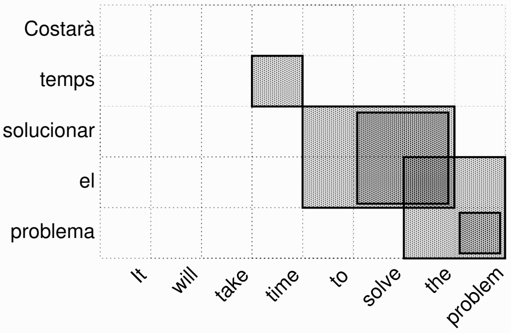
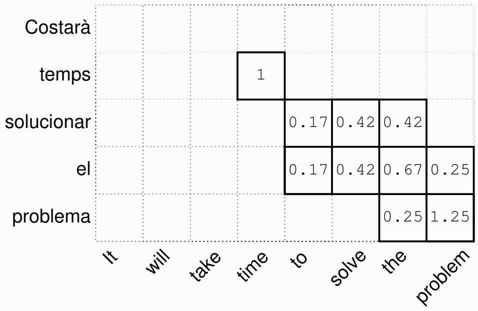


Figure 1: Sub-segment alignments.

Word alignment from sub-segment alignments. The information provided by the SSAs can then be used for word alignment. We define the *alignment strength* A_{jk} between the j-th word in S and the k-th word in T as

$$A_{jk}(S,T,M) = \sum_{(\sigma,\tau)\in M} \frac{\text{cover}(j,k,\sigma,\tau)}{|\sigma|\cdot|\tau|}$$

where M is the set of SSAs detected for the pair of parallel segments S and T, $|x|$ is the length of segment x measured in words, and $\text{cover}(j,k,\sigma,\tau)$ equals 1 if σ covers the j-th word in S and τ the k-th word in T, and 0 otherwise. This way of computing the alignment strengths is based on the idea that SSAs apply *alignment pressures* on the words; so the larger the surface covered by the SSA, the weaker the word-alignment strength obtained. Following our example, the alignment strengths for the words covered by the SSAs are presented in Figure 2. The words *temps* and *time* are only covered by a SSA (*temps*,*time*), so the surface is 1 and the alignment strength is $A_{1,4} = 1$. However, words *the* and *el* are covered by three SSAs: (*solucionar el*, *solve the*), (*solucionar el*, *to solve the*), and (*el problema*, *the problem*). So the alignment strength is $A_{3,6} = 1/4 + 1/6 + 1/4 = 2/3$.



	It	will	take	time	to	solve	the	problem
Costarà								
temps				1				
solucionar					0.17	0.42	0.42	
el					0.17	0.42	0.67	0.25
problema							0.25	1.25

Figure 2: Alignment strengths.

The alignment strengths are then used to obtain word alignments. We simply align the j-th word in S with the k-th word in T if $A_{jk} > 0 \wedge A_{jk} \geq A_{jl}, \forall l \in [1,|T|]$, and vice versa. Note that one word in one of the segments can be aligned with multiple words in the other segment. Figures 3 and 4 show, respectively, the Catalan-to-English and the English-to-Catalan word alignments for the running example.

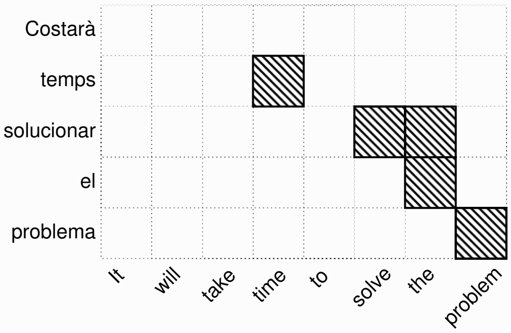


Figure 3: Resulting Catalan to English word alignment.

Figure 5 shows two possible symmetrised word alignments obtained by computing, in the first case, the intersection of the alignments shown in Figures 3 and 4, and, in the second case, the the widely-used *grow-diag-final-and* heuristic (Koehn, Och, and Marcu, 2003). It is worth noting that some words remain unaligned in Figure 5. This is a situation which can also be found in other state-of-the-art word alignment methods and, in this case, can be caused both by the symetrisation method, such as the word *to* in the alignment symetrysed through the intersection, or

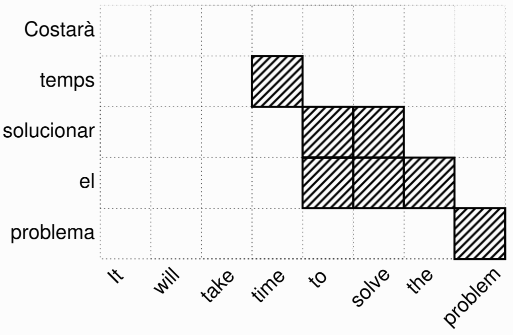


Figure 4: Resulting English to Catalan word alignment.

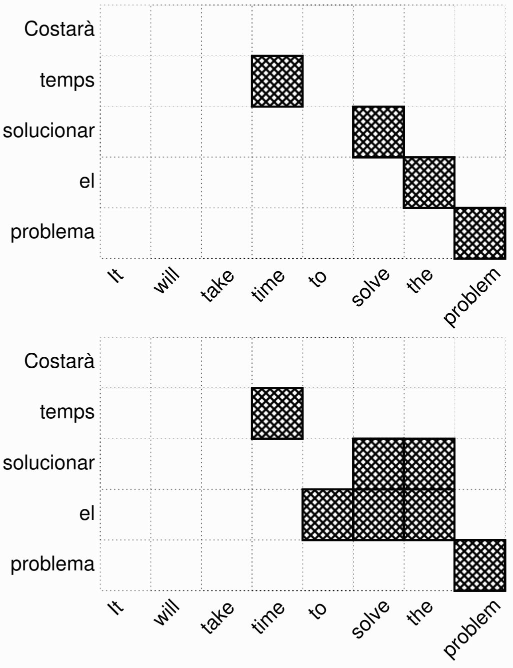


Figure 5: Two possible symmetrised word alignments, the first one using the intersection heuristic and the second one using the grow-diag-final-and heuristic.

by the lack of bilingual evidence relating the words, such as the words *Costarà*, *It*, *will*, and *take*. Depending on the needs of the task, more bilingual sources can be used in order to reduce the number of unaligned words. However, it is worth noting that unaligned words can also be caused by incorrect or excessively free translations, so keeping them unaligned may improve the overall alignment quality.

In addition, alignment strengths can be seen as a measure of the confidence on the relationships between the words. In future works, we plan to use the average alignment strength as a measure of the confidence on the SSAs. In this way, it could be possible to set a threshold to discard less-trusted SSAs. In the running example, the average alignment strength for the SSA (*solucionar el*, *to solve the*) is 0.37, whereas for the SSA (*el problema*, *the problem*) the average alignment strength is 0.60. Therefore, we see that (*el problema*, *the problem*) is a more reliable SSA than (*solucionar el*, *to solve the*).

3 Experimental setting

We evaluated the success of our system for word alignment using a *gold-standard* English–Spanish parallel corpus in which word alignments are annotated. We ran our method in both directions (Spanish to English and English to Spanish) and symmetrised the alignment obtained through the *grow-diag-final-and* heuristic (Koehn, Och, and Marcu, 2003) implemented in Moses (Koehn et al., 2007). We compared the performance of our system with that obtained by GIZA++ (Och and Ney, 2003) in different scenarios.

Test corpus. We used the test parallel corpus from the *tagged EPPS corpus* (Lambert et al., 2005) as a gold-standard parallel corpus.[7] It consists of 400 pairs of sentences from the English–Spanish Europarl (Koehn, 2005) parallel corpus and is provided with the corresponding gold-standard for word alignment. Two levels of confidence are defined for word alignments in this corpus, based on the judgement of the authors of the gold-standard: *sure* alignments and *possible* (less trusted) alignments.

Sources of bilingual information. We used three different MT systems as SBIs to translate the sub-segments from English into Spanish and vice versa:

- *Apertium*:[8] a free/open-source platform for the development of rule-based MT systems (Forcada et al., 2011). We used the English–Spanish MT system from the project's repository[9] (revision 34706).
- *Google Translate*:[10] an online MT system

[7] `http://gps-tsc.upc.es/veu/LR/epps_ensp_alignref.php3` [last visit: 2nd May 2012]

[8] `http://www.apertium.org` [last visit: 2nd May 2012]

[9] `https://apertium.svn.sourceforge.net/svnroot/apertium/trunk/apertium-en-es/` [last visit: 2nd May 2012]

[10] `http://translate.google.com` [last visit: 2nd May 2012]

by Google Inc (translations performed on 28th April 2012).

- *Microsoft Translator*:[11] an online MT system by Microsoft (translations performed on 27th April 2012).

Metrics. We computed the precision (P) and recall (R) for the alignments obtained both by our approach and by the baseline:

$$P = 100\% \cdot \frac{|\text{WA} \cap \text{GS}|}{|\text{WA}|}$$

$$R = 100\% \cdot \frac{|\text{WA} \cap \text{GS}|}{|\text{GS}|}$$

where WA is the set of alignments obtained and GS is the set of alignments in the gold standard. Then, we combined both measures to obtain the F-measure:

$$F = \frac{2 \cdot P \cdot R}{P + R}$$

These three metrics were computed, only for the sure alignments and also for both sure and possible alignments.

Baseline. We compared the performance of our word-alignment method to that of GIZA++ (Och and Ney, 2003), a toolkit for word alignment which implements different statistical alignment strategies. We run GIZA++ in both directions (source to target and target to source) and then we combine both sets of alignments through the *grow-diag-final-and* heuristic (Koehn, Och, and Marcu, 2003).

GIZA++ is widely used for word-alignment in statistical MT. In this scenario, it is usually trained on the parallel corpus to be aligned. However, it is also possible to use pre-trained models to align new pairs of segments, in order to avoid training a new alignment model for each new alignment task. As our system is aimed at performing word alignment on the fly, we consider that the most adequate scenario to compare our approach with GIZA++ is using pre-trained alignment models to align the test corpus. Therefore, for a better comparison of our method to state-of-the-art techniques, we define two baselines. In the first one, henceforth *basic-GIZA++ baseline*, we train and run GIZA++ on the test corpus. In the second one, henceforth *pre-trained-GIZA++ baseline*, we train GIZA++ on a larger parallel corpus and use the resulting models to align the test corpus. To train the alignment models for the pre-trained-GIZA++ baseline, we used the parallel corpus from the News Commentary corpus distributed for the machine translation task in the Workshop on Machine Translation 2011.[12] This corpus was lowercased, tokenized and cleaned to keep only those parallel segments containing up to 40 words. After this process, we obtained a corpus of 126,419 pairs of segments.

segs.	**sure**			**sure ∪ possible**		
	P	**R**	**F**	**P**	**R**	**F**
100	57.1	59.9	58.5	64.7	47.4	54.7
200	57.5	61.2	59.3	64.9	47.5	54.9
300	59.7	63.6	61.6	67.8	50.1	57.7
400	59.9	64.2	62.0	68.2	50.5	58.0

Table 2: Precision (P), recall (R), and F-measure (F) obtained by the basic-GIZA++ baseline for sure alignments, and for all sure and possible alignments when aligning the gold-standard corpus in portions of 100, 200, 300, and 400 pairs of segments (segs).

4 *Results and discussion*

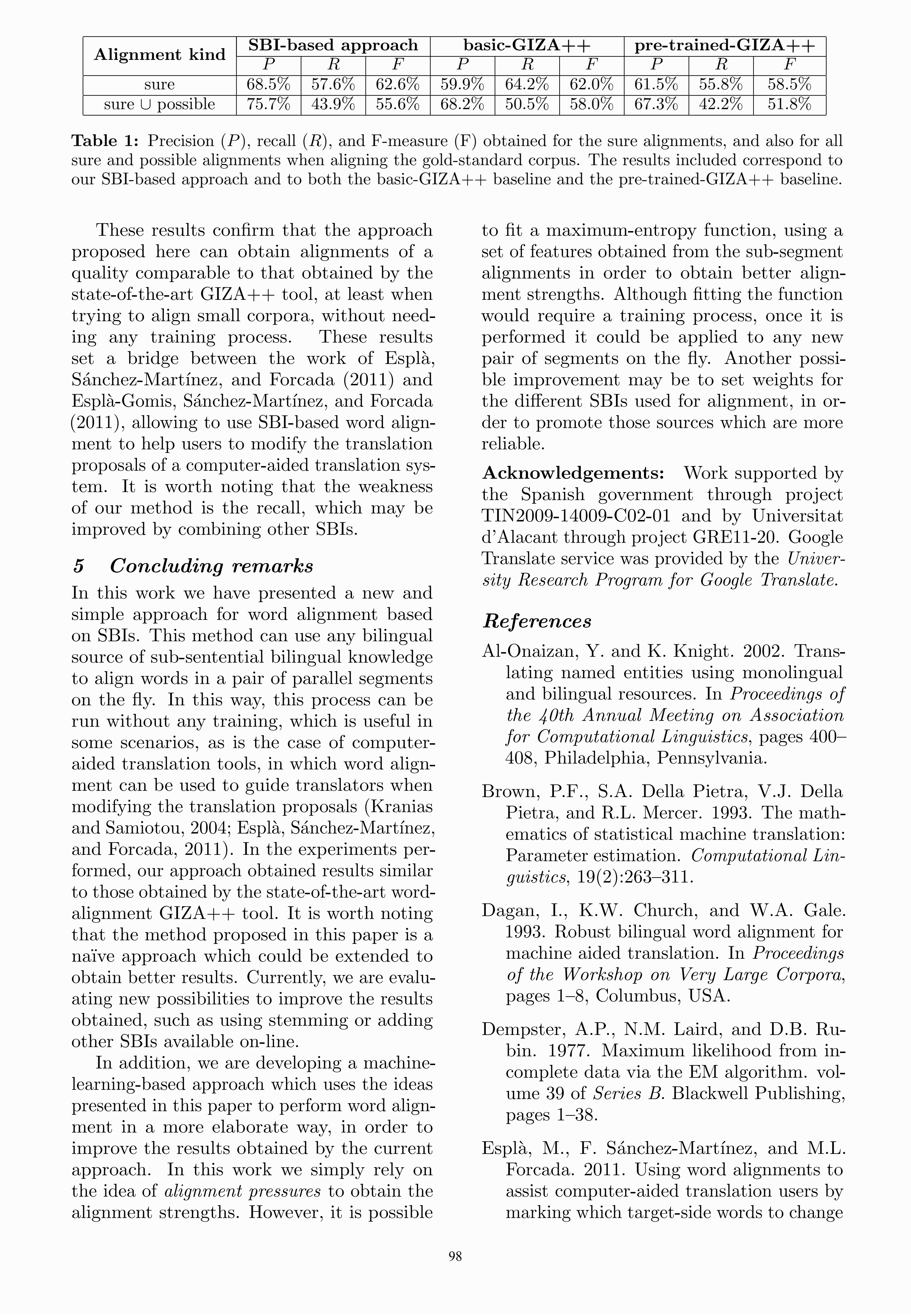
Table 1 shows the results obtained by our system and both baselines based on GIZA++: the basic-GIZA++ baseline and the pre-trained-GIZA++ baseline.

As can be seen, the method proposed in this paper obtains F-measures very similar to those obtained by both GIZA++-based baseline approaches. Another important detail is that our method obtains better precision in alignment than the two baselines proposed, although the results on recall obtained by the basic-GIZA++ baseline are better than ours.

Table 2 presents the results obtained by the basic-GIZA++ baseline when using portions of the test corpus with a different number of pairs of segments. The results presented in this table are useful to understand that, although the basic-GIZA++ yields slightly better results than the other approaches in Table 1, it clearly depends on the size of the parallel corpus to align. Of course, using this approach is not possible when trying to align a pair of segments on the fly, and obtains lower results when trying to align a very small set of parallel segments.

[11] `http://www.microsofttranslator.com` [last visit: 2nd May 2012]

[12] `http://www.statmt.org/wmt11/translation-task.html`



Alignment kind	SBI-based approach			basic-GIZA++			pre-trained-GIZA++		
	P	R	F	P	R	F	P	R	F
sure	68.5%	57.6%	62.6%	59.9%	64.2%	62.0%	61.5%	55.8%	58.5%
sure ∪ possible	75.7%	43.9%	55.6%	68.2%	50.5%	58.0%	67.3%	42.2%	51.8%

Table 1: Precision (P), recall (R), and F-measure (F) obtained for the sure alignments, and also for all sure and possible alignments when aligning the gold-standard corpus. The results included correspond to our SBI-based approach and to both the basic-GIZA++ baseline and the pre-trained-GIZA++ baseline.

These results confirm that the approach proposed here can obtain alignments of a quality comparable to that obtained by the state-of-the-art GIZA++ tool, at least when trying to align small corpora, without needing any training process. These results set a bridge between the work of Esplà, Sánchez-Martínez, and Forcada (2011) and Esplà-Gomis, Sánchez-Martínez, and Forcada (2011), allowing to use SBI-based word alignment to help users to modify the translation proposals of a computer-aided translation system. It is worth noting that the weakness of our method is the recall, which may be improved by combining other SBIs.

5 Concluding remarks

In this work we have presented a new and simple approach for word alignment based on SBIs. This method can use any bilingual source of sub-sentential bilingual knowledge to align words in a pair of parallel segments on the fly. In this way, this process can be run without any training, which is useful in some scenarios, as is the case of computer-aided translation tools, in which word alignment can be used to guide translators when modifying the translation proposals (Kranias and Samiotou, 2004; Esplà, Sánchez-Martínez, and Forcada, 2011). In the experiments performed, our approach obtained results similar to those obtained by the state-of-the-art word-alignment GIZA++ tool. It is worth noting that the method proposed in this paper is a naïve approach which could be extended to obtain better results. Currently, we are evaluating new possibilities to improve the results obtained, such as using stemming or adding other SBIs available on-line.

In addition, we are developing a machine-learning-based approach which uses the ideas presented in this paper to perform word alignment in a more elaborate way, in order to improve the results obtained by the current approach. In this work we simply rely on the idea of *alignment pressures* to obtain the alignment strengths. However, it is possible to fit a maximum-entropy function, using a set of features obtained from the sub-segment alignments in order to obtain better alignment strengths. Although fitting the function would require a training process, once it is performed it could be applied to any new pair of segments on the fly. Another possible improvement may be to set weights for the different SBIs used for alignment, in order to promote those sources which are more reliable.

Acknowledgements: Work supported by the Spanish government through project TIN2009-14009-C02-01 and by Universitat d'Alacant through project GRE11-20. Google Translate service was provided by the *University Research Program for Google Translate.*

References

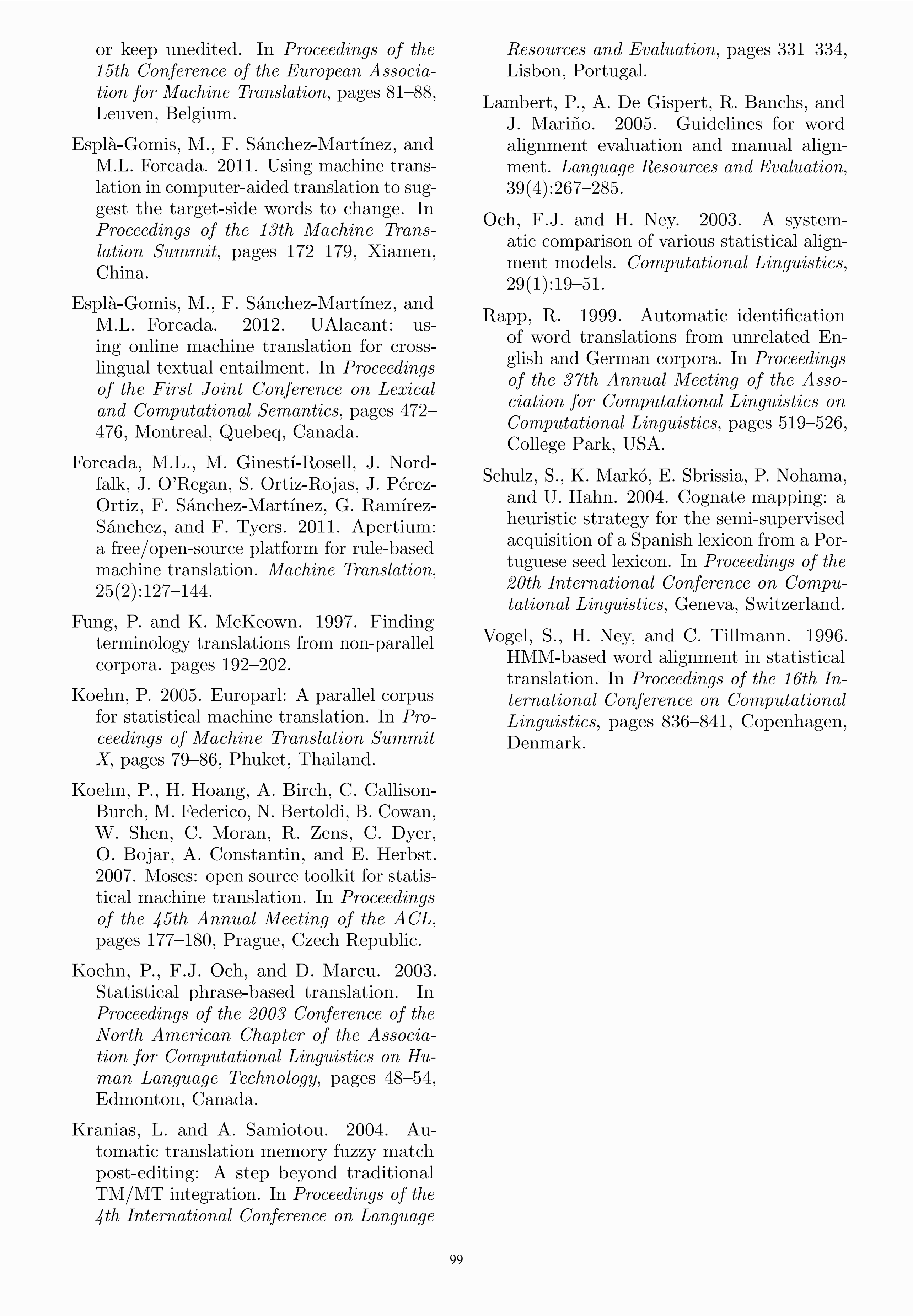
Al-Onaizan, Y. and K. Knight. 2002. Translating named entities using monolingual and bilingual resources. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 400–408, Philadelphia, Pennsylvania.

Brown, P.F., S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Dagan, I., K.W. Church, and W.A. Gale. 1993. Robust bilingual word alignment for machine aided translation. In *Proceedings of the Workshop on Very Large Corpora*, pages 1–8, Columbus, USA.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. volume 39 of *Series B*. Blackwell Publishing, pages 1–38.

Esplà, M., F. Sánchez-Martínez, and M.L. Forcada. 2011. Using word alignments to assist computer-aided translation users by marking which target-side words to change



or keep unedited. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 81–88, Leuven, Belgium.

Esplà-Gomis, M., F. Sánchez-Martínez, and M.L. Forcada. 2011. Using machine translation in computer-aided translation to suggest the target-side words to change. In *Proceedings of the 13th Machine Translation Summit*, pages 172–179, Xiamen, China.

Esplà-Gomis, M., F. Sánchez-Martínez, and M.L. Forcada. 2012. UAlacant: using online machine translation for cross-lingual textual entailment. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 472–476, Montreal, Quebeq, Canada.

Forcada, M.L., M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Fung, P. and K. McKeown. 1997. Finding terminology translations from non-parallel corpora. pages 192–202.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, pages 79–86, Phuket, Thailand.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 177–180, Prague, Czech Republic.

Koehn, P., F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Edmonton, Canada.

Kranias, L. and A. Samiotou. 2004. Automatic translation memory fuzzy match post-editing: A step beyond traditional TM/MT integration. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 331–334, Lisbon, Portugal.

Lambert, P., A. De Gispert, R. Banchs, and J. Mariño. 2005. Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation*, 39(4):267–285.

Och, F.J. and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Rapp, R. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526, College Park, USA.

Schulz, S., K. Markó, E. Sbrissia, P. Nohama, and U. Hahn. 2004. Cognate mapping: a heuristic strategy for the semi-supervised acquisition of a Spanish lexicon from a Portuguese seed lexicon. In *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland.

Vogel, S., H. Ney, and C. Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 836–841, Copenhagen, Denmark.

2.2 Methods for word-level quality estimation in computer-aided translation based on translation memories using external sources of bilingual information

This section describes the research carried out on the use of external SBI to estimate the quality of a translation suggestion at the word level in TM-based CAT. The objective is to confirm that word-keeping recommendations can be provided without the need for a word-alignment intermediate step. Our working hypothesis is:

Hypothesis #3: it is possible to use SBI to estimate the quality of TM-based CAT translation suggestions at the word level.

The research work reported in this section is highlighted and put in context in Figure 2.2 and described in the following article:

- Esplà-Gomis, M., Sánchez-Martínez, F., and Forcada, M.L. 2015. Using machine translation to provide target-language edit hints in computer aided translation based on translation memories. In *Journal of Artificial Intelligence Research*, volume 53, p. 169–222. [**Reprinted publication 2.2.1**]

The approaches described in reprinted publication 2.2.1 build on the ideas and methods developed in Section 2.1 to define a new methodology for word-level QE in TM-based CAT that does not require word alignments because they directly use SBI. There are two such approaches, one heuristic and one based on machine learning techniques, and both are exhaustively evaluated on five different language pairs and using three different SBI as well as texts from different domains. Moreover, these two new approaches are compared to the approaches based on statistical word alignments described in reprinted publication 2.1.1.

It is worth noting that reprinted publication 2.2.1 also includes a study of the productivity gain experienced by professional translators when using word-keeping recommendation in a CAT tool. This study was possible through the development of two plugins for the free/open-source CAT tool OmegaT,[1] one that provides word-keeping recommendations using the heuristic method described in reprinted publication 2.1.1 and colours the words in each translation suggestion (red for the words to be edited, green for the words to be kept untouched);[2] and another that tracks all the actions performed by the translator and the duration of each action.[3] This study, in which five professional translators participated, confirms the relevance of

[1] `http://www.omegat.org`

[2] `http://www.dlsi.ua.es/~mespla/edithints.html`

[3] `https://github.com/mespla/OmegaT-SessionLog`

the problem of word-level QE in TM-based CAT by showing a 14% reduction in the time devoted to a translation task.

The extensive evaluation of the different methods developed in this chapter not only confirms the working hypothesis #3, but also allows to determine which of these approaches are better, depending on the translation task and the resources available.

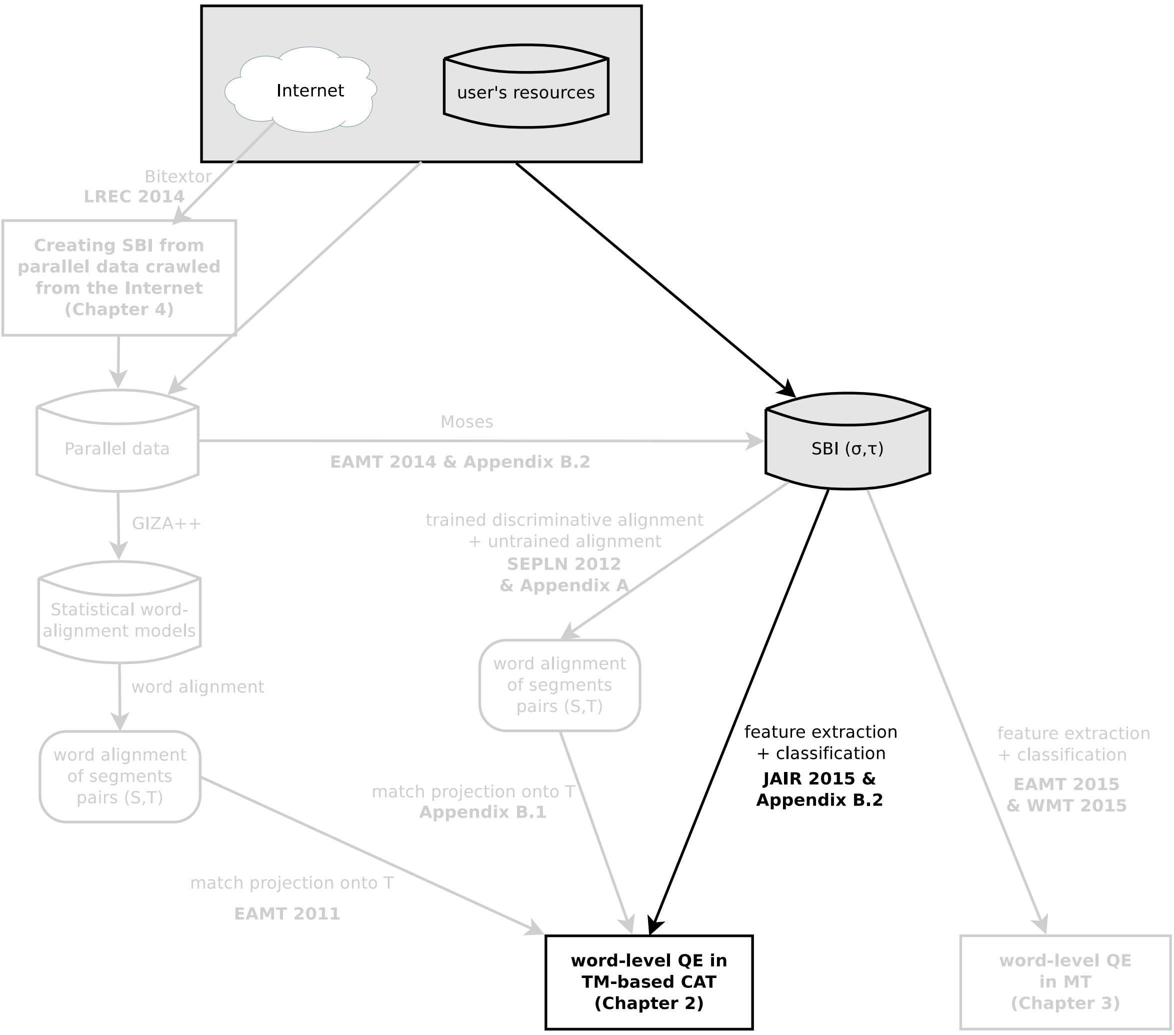

Internet
user's resources
Bitextor
LREC 2014
Creating SBI from parallel data crawled from the Internet (Chapter 4)
Parallel data
Moses
EAMT 2014 & Appendix B.2
SBI (σ,τ)
GIZA++
trained discriminative alignment + untrained alignment
SEPLN 2012 & Appendix A
Statistical word-alignment models
word alignment
word alignment of segments pairs (S,T)
word alignment of segments pairs (S,T)
feature extraction + classification
JAIR 2015 & Appendix B.2
feature extraction + classification
EAMT 2015 & WMT 2015
match projection onto T
Appendix B.1
match projection onto T
EAMT 2011
word-level QE in TM-based CAT (Chapter 2)
word-level QE in MT (Chapter 3)

Figure 2.2: This diagram highlights the research being reported in Section 2.2 (and the publications concerned) on the development of word-level QE methods in TM-based CAT able to directly use external SBI, and relates it to the rest of the work reported in this dissertation.

Using Machine Translation to Provide Target-Language Edit Hints in Computer Aided Translation Based on Translation Memories

Miquel Esplà-Gomis MESPLA@DLSI.UA.ES
Felipe Sánchez-Martínez FSANCHEZ@DLSI.UA.ES
Mikel L. Forcada MLF@DLSI.UA.ES
Dept. de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain

Abstract

This paper explores the use of general-purpose machine translation (MT) in assisting the users of computer-aided translation (CAT) systems based on translation memory (TM) to identify the target words in the translation proposals that need to be changed (either replaced or removed) or kept unedited, a task we term as *word-keeping recommendation*. MT is used as a *black box* to align source and target sub-segments *on the fly* in the translation units (TUs) suggested to the user. Source-language (SL) and target-language (TL) segments in the matching TUs are segmented into overlapping sub-segments of variable length and machine-translated into the TL and the SL, respectively. The bilingual sub-segments obtained and the matching between the SL segment in the TU and the segment to be translated are employed to build the features that are then used by a binary classifier to determine the target words to be changed and those to be kept unedited. In this approach, MT results are never presented to the translator. Two approaches are presented in this work: one using a word-keeping recommendation system which can be trained on the TM used with the CAT system, and a more basic approach which does not require any training.

Experiments are conducted by simulating the translation of texts in several language pairs with corpora belonging to different domains and using three different MT systems. We compare the performance obtained to that of previous works that have used statistical word alignment for word-keeping recommendation, and show that the MT-based approaches presented in this paper are more accurate in most scenarios. In particular, our results confirm that the MT-based approaches are better than the alignment-based approach when using models trained on out-of-domain TMs. Additional experiments were also performed to check how dependent the MT-based recommender is on the language pair and MT system used for training. These experiments confirm a high degree of reusability of the recommendation models across various MT systems, but a low level of reusability across language pairs.

1. Introduction

Computer-aided translation (CAT) systems based on translation memory (TM) (Bowker, 2002; Somers, 2003) are the translation technology of choice for most professional translators, especially when translation tasks are repetitive and the effective recycling of previous translations is feasible. The reasons for this choice are the conceptual simplicity of *fuzzy-match scores* (FMS) (Sikes, 2007) and the ease with which they can be used to determine the usefulness of the translations proposed by the CAT system and to estimate the remain-

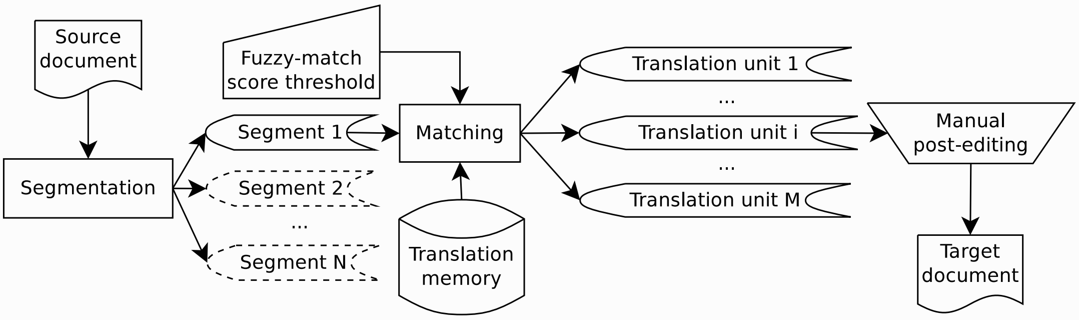


Figure 1: Procedure followed to translate a document using a TM-based CAT system.

ing effort needed to turn them into adequate translations. The FMS function measures the similarity between two text segments, usually by computing a variant of the word-based edit distance (Levenshtein, 1966), although the FMS of some proprietary tools are not publicly described.

When a TM-based CAT system is used to translate a new source document, the system first segments the document, and then, for each source segment S', provides the translator with the subset of translation units (TUs) (S,T) in the TM for which the FMS between S' and S is above a selected threshold Θ. The translator must then choose the TU (S,T) that best fits his or her needs and post-edit its target segment T to produce T', an adequate translation of S'. Figure 1 illustrates this procedure.

When showing the subset of matching TUs to the translator, most TM-based CAT systems highlight the words in S that differ from those in S' in order to ease the task of post-editing T. It is, however, up to the translator to identify the specific words in T that should be changed (either replaced or removed) in order to convert T into T', which is the problem that we deal with in this paper and term as *word-keeping recommendation*. Our experiments with professional translators show that a TM-based CAT system capable of word-keeping recommendation improves their productivity by up to 14% in the ideal case that all recommendations are indeed correct (see Appendix A for more details).

Word-keeping recommendation is related to *translation spotting* (Veronis & Langlais, 2000; Simard, 2003; Sánchez-Martínez, Carrasco, Martínez-Prieto, & Adiego, 2012), which consists of solving the problem of finding parallel sub-segments in parallel texts. Translation spotting is used, for example, by *bilingual concordancers* (Bourdaillet, Huet, Langlais, & Lapalme, 2010), types of tools which help a translator to retrieve occurrences of a sub-segment in a parallel corpus and its corresponding translation. Some examples of commercial bilingual concordancers are *Webitext*,[1] *Linguee*,[2] or *Reverso Context*.[3] Translation spotting is also particularly relevant for example-based machine translation (Somers, 1999), which uses this technique to build the sub-segmental TM used to translate new materials. MT quality estimation (de Gispert, Blackwood, Iglesias, & Byrne, 2013; Specia, Raj, & Turchi, 2010), also shares some features with this task: in both cases the objective is to discover whether

[1] `http://www.webitext.com` [last visit: 15th May 2015]

[2] `http://www.linguee.com` [last visit: 15th May 2015]

[3] `http://context.reverso.net/translation/` [last visit: 15th May 2015]

a translation proposal T^4 is a valid translation for a given source language segment S'. The parallelisms become stronger in the case of word-level quality estimation (Ueffing & Ney, 2005; Bojar et al., 2014; Esplà-Gomis, Sánchez-Martínez, & Forcada, 2015), in which, as in word-keeping recommendation, every word of a proposal is analysed to decide whether or not it is likely to belong to the final translation. There are critical differences between the scenarios in which quality estimation and word-keeping recommendation operate: quality estimation detects words which should be changed in segments T which are likely to be inadequately written in TL, but are intended to be translations of S'; conversely, word-keeping recommendation is intended to work on segments T which are usually adequately written in TL, but they are not a translation of S' (unless an exact match between S and S' is found).

Esplà, Sánchez-Martínez, and Forcada (2011) have performed word-keeping recommendation by using statistical word-alignment models (Och & Ney, 2003) to align the source-language (SL) and target-language (TL) words of each TU in the TM. When a TU (S, T) is suggested to the translator, the pre-computed word alignments are then used to determine the target words to be changed or kept unedited. Analogously, Kranias and Samiotou (2004) align the words in each TU at different sub-segment levels by using, among other resources, a bilingual dictionary of words and phrases (Meyers, Kosaka, & Grishman, 1998), suffix lists to deal with morphological variations, and a list of closed-class words and their categories (Ahrenberg, Andersson, & Merkel, 2000). The authors use these alignments to detect the words to be changed and then use MT to propose a translation for them. To the best of our knowledge, the specific details on how the Kranias and Samiotou method works have not been published. A patent published by Kuhn, Goutte, Isabelle, and Simard (2011) describes a similar method that is also based on statistical word alignment in order to detect the words to be changed in a translation proposal. Unfortunately, the patent does not provide a detailed description of the actual procedure used.

Esplà-Gomis, Sánchez-Martínez, and Forcada (2011) follow a different approach which does not necessitate the computation of word alignments. Instead, they make use of any available MT system as a source of bilingual information to compute a set of features that are used by a perceptron classifier to estimate the probability p_K of each target word being kept unedited. This is done by: obtaining the matching TUs in the TM by using FMS above a given threshold; segmenting the SL and TL segments in each of these TUs into overlapping sub-segments of variable length; machine translating these sub-segments into the TL and the SL, respectively, in order to learn sub-segment alignments; and using these sub-segment alignments and the matching between S and S' to build the features to be used by the classifier. The basic idea behind this method is that a word in T is likely to be kept unedited if it appears in the translation of sub-segments which are common to S and S', the segment to be translated. Finally, p_K is used for word-keeping recommendation by marking the words for which $p_K < \frac{1}{2}$ as "change", or otherwise as "keep".

Although the latter approach requires a training procedure to be run on a TM, Esplà-Gomis et al. (2011) show that, for the translation of Spanish texts into English, the model used by the perceptron classifier can be trained on a TM from a domain that is different from that of the actual TM to be used and the text to be translated. Furthermore, the

[4] In the case of quality estimation, the segment T to be evaluated originates from MT, while in word-keeping recommendation, it originates from a TM-based CAT tool proposal.

results obtained by this system are similar to those obtained by Esplà et al. (2011), based on statistical word alignments, for models trained on texts from the same domain as the text translated, and much better for models trained on out-of-domain texts, as shown in Section 7.

In this paper we revisit the approach of Esplà-Gomis et al. (2011), and propose new feature sets that capture the information in the machine-translated sub-segments in a more successful way than the features therein. In addition, a more complex multilayer perceptron binary classifier is used in this work, which improves the results obtained with the simpler perceptron classifier proposed by Esplà et al. (2011). These improvements on binary classification are compared to the previous approach on a more exhaustive evaluation framework, including new domains and language pairs (see below). Finally, we introduce a new method for word-keeping recommendation that is also able to use any available MT system as a source of bilingual information, and does not require any training procedure to be run in advance. This training-free method uses the sub-segment pairs that match both S and T to compute the *alignment strength* (Esplà-Gomis, Sánchez-Martínez, & Forcada, 2012b) between the words in S and T. The alignment strength between two words s_k in S and t_j in T measures the amount of evidence that relates the two words by giving more weight to the evidence from shorter sub-segments, which involves a sharper picture of the relation between s_k and t_j. Alignment strengths are then used in a similar fashion to that of Esplà et al. (2011) to determine the words to be changed or kept unedited.

As mentioned above, the experiments performed in this work compare the two MT-based approaches (that which requires training and that which is training-free) to the alignment-based approach of Esplà et al. (2011) using ten different language pairs, TMs from three different domains, and three different MT systems. The experiments not only cover the ideal scenario, in which the trained recommendation models are tested under the same conditions —language pair, TM domain, and MT system— used for training, but also scenarios in which some conditions change in order to test the reusability of the models. Namely, experiments were carried out by using recommendation models trained on: a TM from a different domain, a different MT system, and a different language pair. The results obtained show that the MT-based approaches are superior to the alignment-based approach as regards accuracy in all the scenarios. These results additionally confirm that the MT-based approaches produce recommendation models that are more portable across TM domains than those based on word alignment. This also provides good reusability for different MT systems, but poor reusability when translating a different pair of languages to that used for training; in fact, the training-free MT-based method provides better results in this scenario.

The remainder of the paper is organised as follows. The following section reviews the previous works on the integration of TMs and MT. Section 3 reviews the statistical word-alignment-based approach defined by Esplà et al. (2011), which is used in this paper as a reference to compare the new methods presented. Section 4 tackles the problem of word-keeping recommendation by using a binary classifier and several sets of features based on the coverage of sub-segment pairs obtained by machine translating sub-segments of different sizes from the segments in a translation proposal, and the matching between the source-side of the proposal and the segment to be translated. Section 5 then shows how to use these bilingual sub-segments to compute the alignment strength between the SL and TL words in each TU, and how they can be used for word-keeping recommendation without training. Section 6 describes the experimental framework, while Section 7 presents and discusses the results obtained. The paper ends with some concluding remarks. Two appendices are included in this paper: one including experiments aimed at measuring the impact of word-keeping recommendation on the productivity of professional translators, and the other one reporting the results on a filtered-out data set to check their performance in an ideal setting.

2. Integration of Machine Translation and Translation Memories

The literature on this subject contains several approaches that combine the benefits of MT and TMs in ways that are different from those presented in this paper, and that go beyond the obvious *β-combination* scenario defined by Simard and Isabelle (2009), in which MT is used to translate a new segment when no matching TU above a fuzzy-match score threshold β is found in the TM.

Marcu (2001) integrates word-based statistical machine translation (SMT) with a sub-segmental TM. The method uses *IBM model 4* (Brown, Della Pietra, Della Pietra, & Mercer, 1993) word-based translation model to build a sub-segmental TM and learn word-level translation probabilities. This is done by training the the IBM model 4 on the TM used for translation. The source language (SL) segments and the target language (TL) segments in each translation unit (TU) in the TM are then aligned at the word level by using the Viterbi algorithm. Finally, the sub-segmental TM is built from parallel *phrases* in a very similar way to that which occurs in modern phrase-based statistical MT systems (Koehn, 2010): parallel phrases are identified as all those pairs of sub-segments for which all the words on the SL side are aligned with a word on the TL side or with NULL (unaligned), and vice versa. The translation process is then carried out in two stages: first, occurrences of SL phrases in the sub-segmental TM are translated using the corresponding TL phrase; second, words not covered are translated using the word-level translation model learned by the IBM model 4. A similar approach is proposed by Langlais and Simard (2002), who also use translation at sub-segment level. In this case, the segment to be translated is split into sub-segments, and an online bilingual concordancer is used to find their translations. The word-based SMT decoder by Nießen, Vogel, Ney, and Tillmann (1998) is then used to choose the best sub-segments and put them in the best order according to the model.

Biçici and Dymetman (2008) integrate a phrase-based SMT (PBSMT) (Koehn, 2010) system into a TM-based CAT system using discontinuous bilingual sub-segments. The PBSMT system is trained on the same TM, and when a new source segment S' is to be translated, the segments S and T in the best matching TU are used to bias the statistical translation of S' towards T. This is done by augmenting the translation table of the PBSMT system with bilingual sub-segments originating from the fuzzy match (S, T) in which the source part is a common sub-sequence of S and S', and the target part is a sub-sequence of T which has been detected to be aligned with its counterpart sub-sequence in S. Simard and Isabelle (2009) propose a similar approach in which a new feature function is introduced in the linear model combination of a PBSMT system to promote the use of the bilingual sub-segments originating from the best fuzzy match (S, T). Following a similar approach, Läubli, Fishel, Volk, and Weibel (2013) use the mixture-modeling technique (Foster & Kuhn, 2007) to learn a domain-adapted PBSMT system combining an in-domain TM and more general parallel corpora. It is worth noting that none of these three approaches guarantees that the PBSMT system will produce a translation containing the translation in T of the sub-segments that are common to S and S'. In contrast, Zhechev and van Genabith (2010) and Koehn and Senellart (2010), who also use a PBSMT system, do guarantee that the sub-segments of T that have been detected to be aligned with the sub-segments in S matched by S' appear in the translation.

Example-based machine translation (EBMT) (Somers, 1999) has also frequently been used to take advantage of TMs at the sub-segment level (Simard & Langlais, 2001). EBMT systems are based on partial matches from TMs, as in the case of TM CAT tools. In this case, the matching TUs are aligned to detect sub-segment pairs that can be reused for translation. These sub-segment pairs are then combined to produce the most suitable translation for S'. For instance, the commercial TM-based CAT tool *Déjà Vu*[5] integrates example-based MT in order to suggest candidate translations in those cases in which an exact match is not found, but partial matches are available (Garcia, 2005; Lagoudaki, 2008). The example-based-inspired MT system is used to propose a translation by putting together sub-segments of the partial matchings available. Unfortunately, we have been unable to find further details on how this method works.[6] Approaches that combine several MT systems are also available. For example, Gough, Way, and Hearne (2002) use several online MT systems to enlarge the example database of an EBMT system. The authors claim that this permits a better exploitation of the parallel information in the TM for new translations.

Our approach differs from those described above in two ways. First, the aforementioned approaches use the TM to improve the results of MT, or use MT to translate sub-segments of the TUs, while the MT-based approaches presented in this paper use MT to improve the experience of using a TM-based CAT system *without actually showing any machine translated material to the translator.* Second, the approaches above, with the sole exception of that by Gough et al. (2002), focus on a specific MT system or family of MT systems (namely, SMT and EBMT), whereas our MT-based approaches use MT as a black box, and are therefore able to use one or more MT systems at once. In addition, as our MT-based approaches do not need to have access to the inner workings of the MT systems, they are capable of using MT systems that are available on-line (thus avoiding the need for local installation) or even any other source of bilingual information such as dictionaries, glossaries, terminology databases, or sub-segment pairs from bilingual concordancers.

Some works cited in this section, such as those by Zhechev and van Genabith (2010) and Koehn and Senellart (2010) use PBSMT, but they may easily be extended in order to use a different MT system. Their approaches share some similarities with ours: they also try to detect word or phrase alignments as a source of information to find out which parts of a translation proposal should be kept unedited. The main difference between our approach and those by Zhechev and van Genabith and Koehn and Senellart is that they use MT to produce a final translation for the segment to be translated, which comes closer to MT than to TM-based CAT. One of the aims of our approach is to minimally disturb the way translators work with the TM-based CAT system by keeping the translation proposals as found in the TM.

[5] `http://www.atril.com` [last visit: 15th May 2015]

[6] This is usually called "advanced leveraging" (Garcia, 2012).

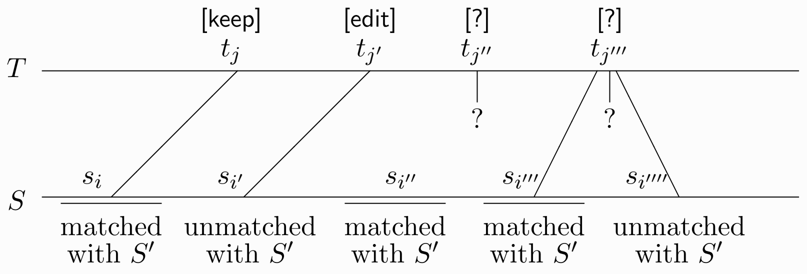


Figure 2: Example of the possible word-alignments that can be obtained for a pair of segments (S, T). Target word t_j may have to remain unedited because it is aligned with source word s_i which is in the part of S that matches S'. Target word $t_{j'}$ may have to be changed because it is aligned with source word $s_{i'}$ which is in the part of S that does not match S'. As target word $t_{j''}$ is not aligned with any source word in S, there is no evidence that could be used to make a recommendation for it. The case of word $t_{j'''}$ is special, since it is aligned with two words, one matching S' and the other not matching S', and a straightforward recommendation cannot be provided.

3. Word-Keeping Recommendation Based on Statistical Word Alignment

This section reviews the first approach for word-keeping recommendation, which was introduced by Esplà et al. (2011), who used statistical word alignment to detect the words to be kept or edited in a translation proposal. Given a segment S' to be translated and a TU (S, T) proposed by the TM-based CAT system, this method first computes the matching between S and S' and aligns the words in S and T by using the word-based statistical translation models implemented in GIZA++ (Och & Ney, 2003). Alignments are then used as follows: let t_j be the word in the j-th position of T which is aligned with a word s_i, the word in the i-th position of S. If s_i is part of the matching between S and S', this indicates that t_j might be part of the translation of S' and that it should therefore remain unedited, as occurs with the word t_j in Figure 2. Conversely, if s_i is not part of the match between S and S', this indicates that t_j might not be the translation of any of the words in S' and it should be edited, as occurs with word $t_{j'}$ in Figure 2. More complex situations in which a TL word is aligned with more than one SL word are tackled by following a voting scheme, as will be explained below. The main limitation of this approach is that, when a word t_j is unaligned, as occurs with the word $t_{j''}$ in Figure 2, there is no evidence that could be used to make a recommendation for it. Although it might be possible to decide on unaligned words by, for example, using the aligned words surrounding them, a wrong recommendation could be worse for the translator than not making any recommendation at all. The idea behind this claim is that a wrong keep recommendation may lead to a wrong translation, which would clearly be undesirable.

In order to determine whether the word t_j in the target proposal T should be changed or kept unedited, the fraction of words aligned with t_j which are common to both S and S'

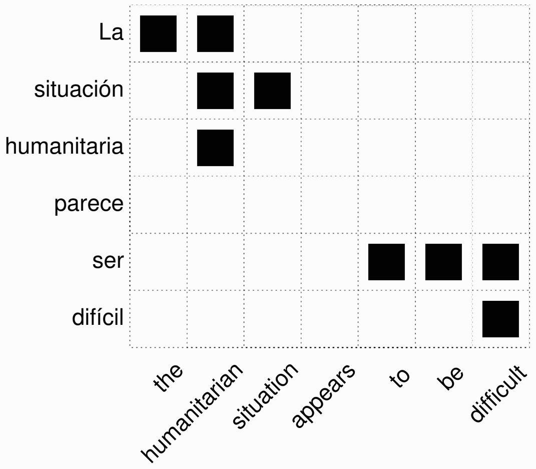


Figure 3: Word alignments for the TU *("la situación humanitaria parece ser difícil", "the humanitarian situation appears to be difficult").*

are computed:

$$f_K(t_j, S', S, T) = \frac{\displaystyle\sum_{s_i \in \text{aligned}(t_j)} \text{matched}(s_i)}{|\text{aligned}(t_j)|}$$

where aligned(t_j) is the set of source words in S which are aligned with target word t_j in T, and matched(s_i) equals 1 if the word s_i is part of the match between S and S', the segment to be translated, and 0 otherwise. Function matched(x) is based on the optimal edit path,[7] obtained as a result of the word-based edit distance (Levenshtein, 1966) between S and S'. The fraction $f_K(t_j, S', S, T)$ may be interpreted as the likelihood of having to keep word t_j unedited. As mentioned above, t_j may be aligned with several words in S, some of which may be common to S and S' while others may not, as occurs with the word $t_{j'''}$ in Figure 2. Esplà et al. (2011) propose two possible heuristics to deal with this:

- *unanimity*: for a word t_j, a recommendation is made if it is aligned only with matched words ($f_K(\cdot) = 1$), or only with unmatched words ($f_K(\cdot) = 0$) in S, while no recommendation is made otherwise; and
- *majority*: this heuristic uses a voting scheme, in which if t_j is aligned with more matched words than unmatched words ($f_K(\cdot) > \frac{1}{2}$), a recommendation is made that it should be kept, and vice versa. Only if t_j is aligned with the same number of matched and unmatched words ($f_K(\cdot) = \frac{1}{2}$) no recommendation is made.

Let us suppose that the TU $(S, T) =$ *("la situación humanitaria parece ser difícil", "the humanitarian situation appears to be difficult")* is proposed in order to translate a new segment $S' =$ *"la situación política parece ser difícil",* and that the word-alignment for

[7] It may occur that more than one optimal path is available to align two segments S and S'. In this case, one of them is chosen arbitrarily.

(S, T) is that depicted in Figure 3. The words *the*, *situation*, *to* and *be* would be marked to be kept, since they are aligned with a single word which is part of the matching between S and S', which is compatible with a possible translation T'= *"the political situation appears to be difficult"*. The word *difficult* would also be marked to be kept, since, even though it is aligned with two words, both are part of the matching between S and S'. However, the evidence for the word *humanitarian* is ambiguous; it is aligned with the words *la* and *situación*, which are part of the matching, but also with *humanitaria* which is not. If the unanimity criterion were to be used, no recommendation would be made for it, while the use of the majority criterion would result in a keeping recommendation. Finally, no recommendation would be made for the word *appears*, since it is not aligned with any word.

The main disadvantage of this approach is that it requires a word-alignment model to be trained directly on the TM to be used for translation in order to maximise the coverage, which means re-training the alignment model every time the TM is updated with new TUs. It may also occur that the TM is not sufficiently large to be able to obtain recommendations with an acceptable quality, signifying that it is necessary to use external parallel corpora in order to train these models. Incremental training (Gao, Lewis, Quirk, & Hwang, 2011) or online training (Bertoldi, Farajian, & Federico, 2009) of statistical word alignment models could be a means to reduce the training time when a TM is modified, or even to adapt a general alignment models to more specific domains, thus improving the coverage. In this case, incremental training would be useful as regards adapting existing models to a new TM, while the on-line training would allow the models to be updated after a new TU has been added to a TM. However, this paper focuses on using machine translation as a source of bilingual information for word-keeping recommendation: we therefore keep the original word-alignment-based approach as described by Esplà et al. (2011) and only use it as a reference when comparing the new approaches proposed here.

4. Word-Keeping Recommendation as Binary Classification

In this work we tackle the problem of word-keeping recommendation as a binary classification problem. For a new segment S' and the TU (S, T) suggested to the translator by the TM-based CAT system, a set of features are computed for each word t_j in T, and a binary classifier is then used to determine whether t_j should be kept unedited or changed (either replaced or deleted). Henceforth, we shall refer to this approach as the *trained MT-based recommender*, to differentiate it from the *training-free MT-based recommender* presented in Section 5.

The features we use are based on the assumption that MT, or any other source of bilingual information, can provide evidence as to whether each word t_j in T should be changed or kept unedited. Let σ be a sub-segment of S from one of the matching TUs (S, T), which is *related by MT* to a sub-segment τ of T. By *related by MT* we mean that machine translating σ leads to τ, or vice versa. We hypothesise that:

- words in σ matching the new segment to be translated S' provide evidence that the words in τ should be kept unedited ("keeping evidence"); and
- words in σ not matching the new segment to be translated S' provide evidence that the words in τ should be changed ("changing evidence").

("la","the") ***["keeping evidence"],***
("situación","situation") ***["keeping evidence"],***
("humanitaria", "humanitarian") ***["changing evidence"],***
("ser","be") ***["keeping evidence"],***
("ser","to be") ***["keeping evidence"],***
("difícil","difficult") ***["keeping evidence"],***
("situación humanitaria","humanitarian situation"),
("ser difícil", "be difficult") ***["keeping evidence"],*** and
("la situación humanitaria", "the humanitarian situation").

Figure 4: Example of the collection M of overlapping machine translated pairs of sub-segments for (S,T) = ("la situación humanitaria parece ser difícil", "the humanitarian situation appears to be difficult"). Sub-segment pairs (σ,τ) with σ matching S and τ matching T are highlighted in bold type.

To continue with the example proposed in Section 3, in which: $(S,T) =$ *("la situación humanitaria parece ser difícil", "the humanitarian situation appears to be difficult"),* and $S'=$ *"la situación política parece ser difícil",* we segment S and T into all possible overlapping sub-segments and translate them with an MT system to obtain the collection M of sub-segment pairs (σ,τ) matching (S,T) and shown in Figure 4. Some sub-segment pairs, such as *("parece", "appear")*, are not included in that list because the translations of the sub-segment on one side do not match their equivalents on the other side. For example, *parece* is translated into English as *seems*, while *appear* is translated into Spanish as *aparecer*. Those pairs (σ,τ) in that list for which all the words in σ match S' provide strong evidence that the words in the corresponding target part should be kept unedited. In this example, these words are *the*, *situation*, *be* and *difficult*, which are compatible with a possible translation $T'=$*"the political situation appears to be difficult"*. Conversely, the pairs (σ,τ) for which all the words in σ do not match S' provide strong evidence that the words in the target part should be changed. In this case, this only occurs for the word *humanitarian*. On the other hand, there is one word about which no evidence can be obtained (*appears*) because it is not matched by the MT system. In this case, it is not possible to provide the translator with any recommendations as occurs, for analogous reasons, with the alignment-based approach described in Section 3. Note that the pair $(\sigma,\tau)=$*("situación humanitaria","humanitarian situation")* contains a source word (*situación*) which matches S' and another (*humanitaria*) which does not match S'. Dealing with this ambiguous evidence, along with combining the evidence from different (σ,τ) (which may be contradictory) leads to an additional problem. In order to deal with ambiguous evidence, we define three feature sets that model and combine "keeping" and "changing" evidence, and which will be described in Sections 4.1, 4.2, and 4.3.

It is worth noting that some pre-processing methods could be used in order to, hopefully, exploit the evidence from bilingual sources of information more efficiently, such as stemming/lemmatisation, morphological analysis, or even the integration of syntactic features such as those proposed by Ma, He, Way, and van Genabith (2011). However, the objective of this approach is to avoid any complex processing in order to obtain fast recommendations when translating texts between any pair of languages, in any domain, and only re-using already available sources of bilingual information, such as the numerous MT systems available on the Internet.

4.1 Features Based on Matching/Mismatching Sub-segments with Unconstrained Length Relations [MM-U]

This feature set was proposed by Esplà-Gomis et al. (2011) and is used as a reference for the remaining feature sets proposed in this work. This feature set considers, for a given (σ, τ) pair of segments, that:

- if σ is a common sub-segment of both the new segment to be translated S' and the source segment S, then it is likely that the words in τ will not have to be changed ("keeping evidence");
- if σ is a sub-segment of S but not of S', then it is likely that the words in τ will have to be changed ("changing evidence").

As can be seen, this is a rather conservative criterion which discards the information from matching words in a partially matching sub-segment σ between S' and S,[8] and will probably be capable of providing high accuracy when recommending that a word should be kept. A more flexible approach is presented in Section 4.2.

Based on the proposed rationale, four sets of features are computed: two sets of *keeping features*, which provide information about the chances of keeping t_j, and two sets of *changing features*, which provide information about the chances of changing t_j. Given the maximum sub-segment length L, the keeping feature set K_{m*} is defined for the word t_j and for every value of $m \in [1, L]$ as follows:

$$K_{m*}(j, S', S, T) = \frac{\text{tcover}(j, \text{seg}_m(S) \cap \text{seg}_m(S'), \text{seg}_*(T), M)}{\text{tcover}(j, \text{seg}_m(S), \text{seg}_*(T), M)},$$

where $\text{seg}_m(X)$ represents the set of all possible m-word sub-segments of segment X, $\text{seg}_*(X)$ is similar to $\text{seg}_m(X)$ but without length constraints, and $\text{tcover}(j, \mathcal{S}, \mathcal{T}, M)$ is defined as:

$$\text{tcover}(j, \mathcal{S}, \mathcal{T}, M) = |\{\tau \in \mathcal{T} : \exists\sigma \in \mathcal{S} \wedge (\sigma, \tau) \in M \wedge j \in \text{span}(\tau, T)\}|,$$

where $\mathcal{S} \subseteq \text{seg}_*(S)$, $\mathcal{T} \subseteq \text{seg}_*(T)$, and function $\text{span}(\tau, T)$ returns the set of word positions spanned by the sub-segment τ in the segment T.[9] Function $\text{tcover}(j, \mathcal{S}, \mathcal{T}, M)$ therefore computes the number of target sub-segments $\tau \in \mathcal{T}$ containing the word t_j that are related by MT to a sub-segment $\sigma \in \mathcal{S}$.

Similarly to K_{m*}, K_{*n} is computed by using only target sub-segments τ of length n:

$$K_{*n}(j, S', S, T) = \frac{\text{tcover}(j, \text{seg}_*(S) \cap \text{seg}_*(S'), \text{seg}_n(T), M)}{\text{tcover}(j, \text{seg}_*(S), \text{seg}_n(T), M)}.$$

[8] For example, a sub-segment of 5 words in which 4 of them are matched and only one is unmatched would be considered as "changing evidence".

[9] Note that a sub-segment τ may be found more than once in segment T: function $\text{span}(\tau, T)$ returns all the possible positions spanned.

Analogously, changing feature sets C_{m*} and C_{*n} are defined as:

$$C_{m*}(j,S',S,T) = \frac{\text{tcover}(j, \text{seg}_m(S) - \text{seg}_m(S'), \text{seg}_*(T), M)}{\text{tcover}(j, \text{seg}_m(S), \text{seg}_*(T), M)},$$

$$C_{*n}(j,S',S,T) = \frac{\text{tcover}(j, \text{seg}_*(S) - \text{seg}_*(S'), \text{seg}_n(T), M)}{\text{tcover}(j, \text{seg}_*(S), \text{seg}_n(T), M)}.$$

In the case of all four features, when both the numerator and the denominator happen to be zero because no pair (σ, τ) covers t_j, the value of the feature is set to $\frac{1}{2}$.

These four features are computed for every value of $1 \leq m \leq L$ and $1 \leq n \leq L$, where L is the maximum sub-segment length used, resulting in $4L$ features. All these features take values in $[0, 1]$ and may have a probabilistic interpretation, in which $\frac{1}{2}$ means "don't know". This feature set will from here on be termed as *MM-U* features. A similar collection of features was tried that constrained the length of both σ (m) and τ (n). However, the results confirmed that no improvement was obtained by adding it to the feature set.

For the running example, we show the feature set that could be computed at word *be*, the sixth word in T (t_6). Please recall that (S, T) = *("la situación humanitaria parece ser difícil", "the humanitarian situation appears to be difficult").* Using the collection of translated pairs of overlapping sub-segments shown in Figure 4, there are three sub-segment pairs (σ, τ) in M that cover the word *be*:

$$\boldsymbol{M} = \{(\text{“}\textbf{ser}\text{”}, \text{“}\textbf{be}\text{”}), (\text{“}\textbf{ser}\text{”}, \text{“}\textbf{to be}\text{”}), (\text{“}\textbf{ser difícil}\text{”}, \text{“}\textbf{be difficult}\text{”})\}.$$

These pairs (σ, τ) only contain sub-segments σ with $m \in [1, 2]$. The value of the function tcover is:

$$\text{tcover}(6, \text{seg}_1(S), \text{seg}_*(T), M) = |\{\text{“be”}, \text{“to be”}\}| = 2$$
$$\text{tcover}(6, \text{seg}_2(S), \text{seg}_*(T), M) = |\{\text{“be difficult”}\}| = 1$$

for both values of m. In addition, the value of tcover for all the sub-segments σ that match S' is:

$$\text{tcover}(6, \text{seg}_1(S) \cap \text{seg}_1(S'), \text{seg}_*(T), M) = |\{\text{“be”}, \text{“to be”}\}| = 2$$
$$\text{tcover}(6, \text{seg}_2(S) \cap \text{seg}_2(S'), \text{seg}_*(T), M) = |\{\text{“be difficult”}\}| = 1$$

The value of the corresponding features is therefore:

$$K_{1*}(6, S', S, T) = \frac{\text{tcover}(6, \text{seg}_1(S) \cap \text{seg}_1(S'), \text{seg}_*(T), M)}{\text{tcover}(6, \text{seg}_1(S), \text{seg}_*(T), M)} = \frac{2}{2} = 1$$

$$K_{2*}(6, S', S, T) = \frac{\text{tcover}(6, \text{seg}_2(S) \cap \text{seg}_2(S'), \text{seg}_*(T), M)}{\text{tcover}(6, \text{seg}_2(S), \text{seg}_*(T), M)} = \frac{1}{1} = 1$$

Features $K_{*1}(6, S', S, T)$ and $K_{*2}(6, S', S, T)$ can be computed analogously. This case is rather simple, since all the evidence available for this word indicates that it should be kept. However, for the word *situation* (t_3), both "keeping" and "changing" evidence coexist in the set M of translated sub-segments pairs:

$$\boldsymbol{M} = \{(\text{“}\textbf{situación}\text{”}, \text{“}\textbf{situation}\text{”}), (\text{“situación humanitaria”}, \text{“humanitarian situation”}),$$
$$(\text{“la situación humanitaria”}, \text{“the humanitarian situation”})\}$$

In this case, τ sub-segments take lengths $m \in [1,3]$, which produces the following values for tcover$(\cdot)$:

$$\text{tcover}(3, \text{seg}_1(S), \text{seg}_*(T), M) = |\{\text{“situation”}\}| = 1$$

$$\text{tcover}(3, \text{seg}_2(S), \text{seg}_*(T), M) = |\{\text{“humanitarian situation”}\}| = 1$$

$$\text{tcover}(3, \text{seg}_3(S), \text{seg}_*(T), M) = |\{\text{“the humanitarian situation”}\}| = 1$$

However, in this case, not all of them match S':

$$\text{tcover}(3, \text{seg}_1(S) \cap \text{seg}_1(S'), \text{seg}_*(T), M) = |\{\text{“situation”}\}| = 1$$

$$\text{tcover}(3, \text{seg}_2(S) \cap \text{seg}_2(S'), \text{seg}_*(T), M) = |\emptyset| = 0$$

$$\text{tcover}(3, \text{seg}_3(S) \cap \text{seg}_3(S'), \text{seg}_*(T), M) = |\emptyset| = 0$$

This allows us to compute the following keeping features:

$$K_{1*}(3, S', S, T) = \frac{\text{tcover}(3, \text{seg}_1(S) \cap \text{seg}_1(S'), \text{seg}_*(T), M)}{\text{tcover}(3, \text{seg}_1(S), \text{seg}_*(T), M)} = \frac{1}{1} = 1$$

$$K_{2*}(3, S', S, T) = \frac{\text{tcover}(3, \text{seg}_2(S) \cap \text{seg}_2(S'), \text{seg}_*(T), M)}{\text{tcover}(3, \text{seg}_2(S), \text{seg}_*(T), M)} = \frac{0}{1} = 0$$

$$K_{3*}(3, S', S, T) = \frac{\text{tcover}(3, \text{seg}_3(S) \cap \text{seg}_3(S'), \text{seg}_*(T), M)}{\text{tcover}(3, \text{seg}_3(S), \text{seg}_*(T), M)} = \frac{0}{1} = 0$$

Analogously, for the changing features, we have:

$$\text{tcover}(3, \text{seg}_1(S) - \text{seg}_1(S'), \text{seg}_*(T), M) = |\emptyset| = 0$$

$$\text{tcover}(3, \text{seg}_2(S) - \text{seg}_2(S'), \text{seg}_*(T), M) = |\{\text{“humanitarian situation”}\}| = 1$$

$$\text{tcover}(3, \text{seg}_3(S) - \text{seg}_3(S'), \text{seg}_*(T), M) = |\{\text{“the humanitarian situation”}\}| = 1$$

which allow us to obtain the following features:

$$C_{1*}(3, S', S, T) = \frac{\text{tcover}(3, \text{seg}_1(S) - \text{seg}_1(S'), \text{seg}_*(T), M)}{\text{tcover}(3, \text{seg}_1(S), \text{seg}_*(T), M)} = \frac{0}{1} = 0$$

$$C_{2*}(3, S', S, T) = \frac{\text{tcover}(3, \text{seg}_2(S) - \text{seg}_2(S'), \text{seg}_*(T), M)}{\text{tcover}(3, \text{seg}_2(S), \text{seg}_*(T), M)} = \frac{1}{1} = 1$$

$$C_{3*}(3, S', S, T) = \frac{\text{tcover}(3, \text{seg}_3(S) - \text{seg}_3(S'), \text{seg}_*(T), M)}{\text{tcover}(3, \text{seg}_3(S), \text{seg}_*(T), M)} = \frac{1}{1} = 1$$

In this case, the ambiguity in the features will be managed by the binary classifier, which will determine the corresponding weights during training.

4.2 Features Based on Partially Matching Sub-segments with Constrained Length Relations [PM-C]

This feature set is slightly different from the previous one as regards the way in which the evidence from the pairs of sub-segments $(\sigma, \tau) \in M$ is used. In this case, the features represent the fraction of words in S that match S' to which a given word t_j in T is related by means of sub-segment pairs (σ, τ). It is worth noting that in the previous feature set, the matching of sub-segment pairs (σ, τ) was evaluated for the whole sub-segment σ. However, in this new feature set, both the keeping and the changing features are computed using the matched/unmatched words in σ. The objective of this feature set is to use the positive evidence from partially matching sub-segments σ more efficiently. The following equation defines the new keeping feature K^W_{mn}:

$$K^W_{mn}(j, S', S, T) = \sum_{k=1}^{|S|} \text{stcover}(j, k, \text{seg}_m(S), \text{seg}_n(T), M) \times \text{match}(k, S', S)$$

where j is the position of t_j in T, k is the position of s_k in S, $\text{match}(k, S', S)$ is 1 if s_k is part of the match between S and S', and 0 otherwise,[10] and function $\text{stcover}(j, k, \mathcal{S}, \mathcal{T}, M)$ is defined as:

$$\text{stcover}(j, k, \mathcal{S}, \mathcal{T}, M) = |\{(\sigma, \tau) \in M : \tau \in \mathcal{T} \wedge \sigma \in \mathcal{S} \wedge j \in \text{span}(\tau, T) \wedge k \in \text{span}(\sigma, S)\}|$$

Similarly, we define the *changing feature* C^W_{mn} as:

$$C^W_{mn}(j, S', S, T) = \sum_{k=1}^{|S|} \text{stcover}(j, k, \text{seg}_m(S), \text{seg}_n(T), M) \times (1 - \text{match}(k, S', S)).$$

Function $\text{stcover}(j, k, \mathcal{S}, \mathcal{T}, M)$ differs from $\text{tcover}(j, \mathcal{S}, \mathcal{T}, M)$ in that, for a given pair (σ, τ), the former takes into account both σ and τ while the latter only takes into account τ. This makes K^W_{mn} and C^W_{mn} complementary, whereas K_{mn} and C_{mn} are not. K^W_{mn} and C^W_{mn} may be combined in a single normalised feature that we term as KC^W_{mn}:

$$\text{KC}^W_{mn}(j, S', S, T) = \frac{\displaystyle\sum_{k=1}^{|S|} \text{stcover}(j, k, \text{seg}_m(S), \text{seg}_n(T), M) \times \text{match}(k, S', S)}{\displaystyle\sum_{k=1}^{|S|} \text{stcover}(j, k, \text{seg}_m(S), \text{seg}_n(T), M)}, \qquad (1)$$

As in the feature set described in Section 4.1, KC^W_{mn} takes values in $[0, 1]$, and, as in that case, when no evidence is found for t_j, the value of the corresponding feature is set to $\frac{1}{2}$. This feature set results in L^2 features and will be referred to as *PM-C*.

[10] The function $\text{match}(k, S', S)$ is based on the optimal edit path obtained as a result of the word-based edit distance (Levenshtein, 1966) between S' and S. Although this is not frequent, it may occur that more than one optimal paths are available: in this case, one of them is chosen arbitrarily.

For the running example, we compute the PM-C features for the word *situation*, as occurred in Section 4.1. As in the previous example, we use the collection of translated sub-segments in Figure 4. The set of sub-segments pairs covering the word *situation* is:

$$\boldsymbol{M} = \boldsymbol{\{}(\text{“}\textbf{situación}\text{”}, \text{“}\textbf{situation}\text{”}), (\text{“situación humanitaria”}, \text{“humanitarian situation”}), \\ (\text{“la situación humanitaria”}, \text{“the humanitarian situation”})\boldsymbol{\}}$$

and the features $KC^W_{1,1}(3, S', S, T)$, $KC^W_{2,2}(3, S', S, T)$, and $KC^W_{3,3}(3, S', S, T)$ can be computed from them. As can be seen stcover$(\cdot)$ happens to be different to zero only for $k = 1$:

$$\text{stcover}(3, 1, \text{seg}_1(S), \text{seg}_1(T), M) = |\{(\text{“}\textbf{situación}\text{”}, \text{“situation”})\}| = 1$$

In this case, we see that *situación* (s_2) is related to *situation* through the sub-segment pair (σ, τ)=(“situación”, “situation”). In this case, σ completely matches S', and we therefore have that:

$$KC^W_{1,1}(3, S', S, T) = \frac{\sum_{k=1}^{|S|} \text{stcover}(3, k, \text{seg}_1(S), \text{seg}_1(T), M) \times \text{match}(k, S', S)}{\sum_{k=1}^{|S|} \text{stcover}(3, k, \text{seg}_1(S), \text{seg}_1(T), M)} = \frac{1}{1} = 1$$

The case of $KC^W_{2,2}(3, S', S, T)$ is slightly more complex. Here, stcover$(\cdot)$ happens to be different to zero only for $k \in [1, 2]$:

$$\text{stcover}(3, 1, \text{seg}_2(S), \text{seg}_2(T), M) = \\ |\{(\text{“}\textbf{situación}\text{ humanitaria”}, \text{“humanitarian situation”})\}| = 1$$

$$\text{stcover}(3, 2, \text{seg}_2(S), \text{seg}_2(T), M) = \\ |\{(\text{“situación }\textbf{humanitaria}\text{”}, \text{“humanitarian situation”})\}| = 1$$

As will be observed, the words *situación* and *humanitaria* are related to *situation* through the same pair (σ, τ)=(“situación humanitaria”,“humanitarian situation”). Here, only one of the two words matches S', hence:

$$KC^W_{2,2}(3, S', S, T) = \frac{\sum_{k=1}^{|S|} \text{stcover}(3, k, \text{seg}_2(S), \text{seg}_2(T), M) \times \text{match}(k, S', S)}{\sum_{k=1}^{|S|} \text{stcover}(3, k, \text{seg}_2(S), \text{seg}_2(T), M)} = \frac{1}{2} = 0.5$$

Finally we have that, for $KC^W_{3,3}(3, S', S, T)$, stcover$(\cdot)$ happens to be different to zero for $k \in [1, 3]$:

$$\text{stcover}(3, 1, \text{seg}_3(S), \text{seg}_3(T), M) = \\ |\{(\text{“}\textbf{la}\text{ situación humanitaria”}, \text{“the humanitarian situation”})\}| = 1$$

$$\text{stcover}(3, 2, \text{seg}_3(S), \text{seg}_3(T), M) = \\ |\{(\text{“la } \textbf{situación} \text{ humanitaria”}, \text{“the humanitarian situation”})\}| = 1$$

$$\text{stcover}(3, 3, \text{seg}_3(S), \text{seg}_3(T), M) = \\ |\{(\text{“la situación } \textbf{humanitaria}\text{”}, \text{“the humanitarian situation”})\}| = 1$$

This time, three words are related to *situation*, all of them through the same sub-segment pair (σ, τ)=(“la situación humanitaria”,“the humanitarian situation”). In this case, *la* and *situación* match S', while *humanitaria* does not. The resulting feature is therefore:

$$KC^W_{3,3}(3, S', S, T) = \frac{\displaystyle\sum_{k=1}^{|S|} \text{stcover}(3, k, \text{seg}_3(S), \text{seg}_3(T), M) \times \text{match}(k, S', S)}{\displaystyle\sum_{k=1}^{|S|} \text{stcover}(3, k, \text{seg}_3(S), \text{seg}_3(T), M)} = \frac{2}{3} \simeq 0.67$$

Note that this feature collection constrains the length of σ and τ at the same time. This configuration was also tried in the previous feature set (MM-U), but no improvements were obtained as compared to constraining the lengths of σ and τ separately. With this feature set both possibilities were also tried, but here constraining the length of σ and τ at the same time proved to lead to better results.

4.3 Features Combining Partially Matching Sub-segments with Constrained Length Relations and Information about Coverage [PM-C+C]

Features KC^W_{mn} above may hide the amount of “keeping/changing evidence”, since they only take into account the fraction of “keeping evidence” from the total amount of evidence.[11] To deal with this, we propose that the feature set defined in Section 4.2 be combined with a new feature E_{mn}:

$$E_{mn}(j, S, T) = |\{(\sigma, \tau) \in M : \sigma \in \text{seg}_m(S) \wedge \tau \in \text{seg}_n(T) \wedge j \in \text{span}(\tau, T)\}| \quad (2)$$

This feature counts the number of sub-segment pairs (σ, τ) covering word t_j, thus providing a measure of the amount of evidence supporting the value of feature KC^W_{mn}. We propose a new feature set, with $2L^2$ features, using both KC^W_{mn} and $E_{mn}(j, S', S, T)$, which will be referred to as *PM-C+C*. A similar feature set was tried, in which E_{mn} was normalised by dividing it by the maximum number of pairs (σ, τ) that could have covered t_j $(m * n)$. However, this set of features did not show any improvement and was therefore discarded.

For the running example, the pairs (σ, τ) in M that cover the word *be* (t_6) are:

$$\boldsymbol{M} = \{(\textbf{“ser”}, \textbf{“be”}), (\textbf{“ser”}, \textbf{“to be”}), (\textbf{“ser difícil”}, \textbf{“be difficult”})\}.$$

Therefore, the features E_{mn} that are different to zero for the word *be* are:

$$E_{1,1}(6, S, T) = |\{(\text{“ser”}, \text{“be”})\}| = 1$$

[11] For example, it would be the same to have 1 “keeping evidence” out of 1 evidence and 5 keeping evidences out of 5 evidences; however, the second case should be considered to be more reliable, since more evidence confirms the keeping recommendation.

$$E_{1,2}(6, S, T) = |\{(\text{“ser”}, \text{“to be”})\}| = 1$$

$$E_{2,2}(6, S, T) = |\{(\text{“ser difícil”}, \text{“be difficult”})\}| = 1$$

Given that a single pair (σ, τ) covers the word *be* with the same m and n, all these features are set to 1. However, if the evidence (σ, τ)=*(“ser”, “be difficult”)* could be used, the value of $E_{1,2}$ would become higher:

$$E_{1,2}(6, S, T) = |\{(\text{“ser”}, \text{“to be”}), (\text{“ser”}, \text{“be difficult”})\}| = 2$$

5. Word-Keeping Recommendation Based on MT Alignment Strengths

The collection M of sub-segment pairs (σ, τ) which are related by MT can be used for word-keeping recommendation directly, i.e., without having to run any training procedure. We propose to do this by using a *training-free* MT-based recommender based on the *alignment strengths* described by Esplà-Gomis, Sánchez-Martínez, and Forcada (2012a) and Esplà-Gomis et al. (2012b). This metric determines the relatedness or association strength between the j-th word in T and the k-th word in S and is defined as:

$$A(j, k, S, T) = \sum_{m=1}^{L} \sum_{n=1}^{L} \frac{\text{stcover}(j, k, \text{seg}_m(S), \text{seg}_n(T), M)}{m \times n}$$

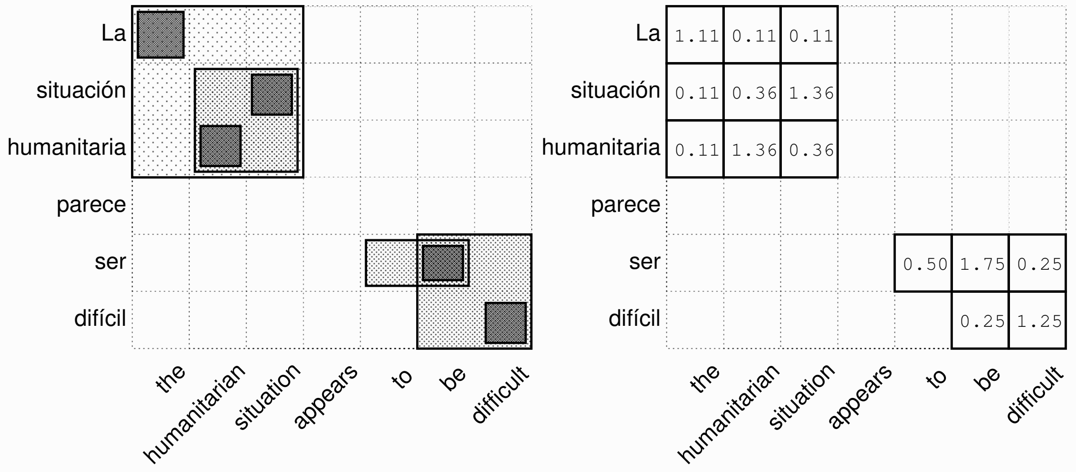
The alignment strength is based on the idea that matched sub-segment pairs apply *pressure* to the words, signifying that the larger the surface covered by a sub-segment pair, the lower the pressure applied to each individual word. Figure 5 shows how the words in a TU are covered by the bilingual sub-segments in M (left), and the result of computing the alignment strengths (right).

In order to perform a word-keeping recommendation using the alignment strengths, we define function $G(j, S', S, T)$ which computes the fraction of the alignment strength that relates a word t_j to those words s_k which are part of the matching between S' and S, over the sum of the alignment strength for all the words in S:

$$G(j, S', S, T) = \frac{\sum_{k=1}^{|S|} A(j, k, S, T) \times \text{match}(k, S', S)}{\sum_{k=1}^{|S|} A(j, k, S, T)} \tag{3}$$

Word keeping recommendation is then performed in the following simple manner: if $G(j, S', S, T) \leq \frac{1}{2}$, then word t_j is marked to be changed, otherwise to be kept. If there is no evidence (σ, τ) with τ spanning t_j, then no recommendation is provided for t_j.

It is worth noting that $A(j, k, S, T)$ is similar to a particular linear combination of the PM-C feature set described in Section 4.2, in which the weight of each feature is directly set to $\frac{1}{mn}$, rather than being chosen by optimising the recommendation accuracy in a training set. The results shown in Section 6 prove that this method is less accurate than the trained MT-based approach, but still provides reasonably good results.



	the	humanitarian	situation	appears	to	be	difficult
La	1.11	0.11	0.11				
situación	0.11	0.36	1.36				
humanitaria	0.11	1.36	0.36				
parece							
ser					0.50	1.75	0.25
difícil						0.25	1.25

Figure 5: Sub-segment pairs covering the words in the TU *("la situación humanitaria parece ser difícil", "the humanitarian situation appears to be difficult")* (left), and the alignment strengths obtained from them (right). The weight of each sub-segment pair is taken to be 1 and is divided by the surface it covers to compute the "pressure" exerted on each individual word.

For the running example, we use the scores shown in Figure 5; as can be seen, the word *situation* is related to three words: *La*, with a score of 0.11, *situación*, with a score of 1.36, and *humanitaria*, with a score of 0.36. The value of $G(3, S', S, T)$ is therefore:

$$G(3, S', S, T) = \frac{0.11 \times 1 + 1.36 \times 1 + 0.36 \times 0}{0.11 + 1.36 + 0.36} = \frac{1.47}{1.83} \simeq 0.8$$

In this case, $G(3, S', S, T) > \frac{1}{2}$, which means that the word *situation* must remain unedited. However, for *humanitarian*, which is related to the same words in Spanish, we have:

$$G(2, S', S, T) = \frac{0.11 \times 1 + 0.36 \times 1 + 1.36 \times 0}{0.11 + 0.36 + 1.36} = \frac{0.47}{1.83} \simeq 0.3$$

Since $G(2, S', S, T) \leq \frac{1}{2}$, the word *humanitarian* would be marked to be changed.

6. Experimental Settings

The experiments conducted consisted of simulating the translation of texts between several language pairs and text domains. The language pairs involved in the experiments were German–English (de→en), English–German (en→de), English–Spanish (en→es), Spanish–English (es→en), English–Finnish (en→fi), Finnish–English (fi→en), English–French (en→fr), French–English (fr→en), Spanish–French (es→fr), and French–Spanish (fr→es). Three thematic TMs were created for each language pair by extracting domain-specific TUs from the DGT-TM (Steinberger, Eisele, Klocek, Pilos, & Schlüter, 2012), a TM published by the *European Commission Directorate-General for Translation* (European Commission Directorate-General for Translation, 2009).

We compared the two MT-based approaches described in Sections 4 and 5 with a *naïve* baseline which is also based on a binary classifier, but using only the fuzzy-match score (FMS) between S and S' as a feature, (henceforth *FMS-only baseline*, see Section 6.5), and with the statistical word-alignment-based approach described in Section 3, in different scenarios. We first evaluated all the approaches in the optimal case, in which the models (either word-alignment models or classification models) are trained on the same TM used for translating, and, in the case of the MT-based approaches, employing the same MT system used for translation. This evaluation was then extended to evaluate:

- *the reusability across domains*: the word-alignment models and the classification models are trained on out-of-domain TMs;
- *the reusability across MT systems*: the models are trained on the same TM used for translation, but using a different MT system; and
- *the reusability across language pairs*: the models are trained on a TM from the same domain as that used for translation, but with a different language pair, and obviously, using a different MT system.

The reusability across MT systems can evidently be evaluated only in the case of MT-based approaches. As regards reusability across language pairs, on the one hand, the FMS-only baseline is language independent, while on the other, the statistical word-alignment models used by the alignment-based approach have to be trained on the same pair of languages as that used for translation.

This extensive evaluation will allow us to ascertain the degree of independence of the recommendation model with regard to the domain of the TM, the MT system, and the language pair used during training. This is a key point, since high independence of all or part of these variables would allow computer-aided translation (CAT) users to reuse existing feature weights obtained without having to run any training procedure when they change the domain of the texts to be translated, the MT system they use or even the languages they are working with. The case of domain independence is particularly relevant since it covers not only the problem of using a different TM, but also the case in which new TUs which were not seen during training are added to a TM.

With regard to the MT systems, we have used the statistical MT system by Google,[12] Power Translator (Depraetere, 2008) version 15,[13] and the free/open-source, shallow-transfer MT system Apertium (Forcada et al., 2011).[14] Unfortunately, not all the MT systems mentioned above were available for all the language pairs. Table 1 shows the MT system(s) available for each language pair included in the experiments.

Even though we used large data sets in a batch mode to obtain the results reported in this paper, we wanted to ensure that the MT-based approaches would be able to provide recommendations in real time for translation tasks. The main part of the computation time for the MT-based approaches is spent segmenting S and T and machine-translating the resulting sub-segments. In order to prove that this could be done in a real MT-based

[12] http://translate.google.com [last visit: 15th May 2015]

[13] http://www.lec.com/power-translator-software.asp [last visit: 15th May 2015]

[14] http://www.apertium.org [last visit: 15th May 2015]

Language pair	Apertium	Google Translate	Power Translator
German→English (`de`→`en`)		✓	✓
English→German (`en`→`de`)		✓	✓
English→Spanish (`en`→`es`)	✓	✓	✓
Spanish→English (`es`→`en`)	✓	✓	✓
English→Finnish (`en`→`fi`)		✓	
Finnish→English (`fi`→`en`)		✓	
English→French (`en`→`fr`)		✓	✓
French→English (`fr`→`en`)		✓	✓
Spanish→French (`es`→`fr`)	✓	✓	✓
French→Spanish (`fr`→`es`)	✓	✓	✓

Table 1: MT systems available for each translation direction (→) used in the experiments.

CAT scenario, a prototype[15] plug-in implementing the training-free approach was built for the free/open-source CAT system OmegaT[16] and, after some experiments using the on-line MT systems Apertium and Google Translate, we can confirm that recommendations are obtained almost instantaneously.

6.1 Evaluation

The FMS-only baseline, the statistical alignment-based approach proposed in Section 3, and the MT-based approaches proposed in Sections 4, and 5 were tested by using a test set (TS) of parallel segments $\{(S'_l, T'_l)\}_{l=1}^{N}$, and a TM in the same domain. For each SL segment S'_l in TS, the set of matching TUs $\{(S_i, T_i)\}_{i=1}^{N'}$ in the TM with a FMS above threshold Θ is obtained. Please recall that the FMS measures the similarity between the translation proposals and the segments to be translated. The FMS threshold Θ is usually set to values above 60% (Bowker, 2002, p. 100) and in our experiments we therefore used several values of Θ of between 60% and 90%.[17] Once the set of matching TUs has been obtained, the recommendations for every word t_j in every target-language segment T_i are obtained and evaluated by using T'_l, the translation of S'_l, as a *gold standard*. The words in T'_l and T_i are matched by using the Levenshtein edit distance (Levenshtein, 1966), which allows us to check whether or not a given word t_j, the j-th word of T_i, is actually kept in the final translation. It is thus possible to determine whether a recommendation for t_j is successful both if:

- t_j is recommended to be changed in T_i and it does not match any word in T'_l, or
- t_j is recommended to be kept in T_i and it does match a word in T'_l.

[15] `http://www.dlsi.ua.es/~mespla/edithints.html` [last visit: 15th May 2015]

[16] `http://www.omegat.org` [last visit: 15th May 2015]

[17] For those MT-based approaches that require training, different models were trained for every value of Θ included in the experiments.

Once all the pairs $(S'_l, T'_l) \in$ TS have been used to obtain their corresponding sets of matching TUs $\{(S_i, T_i)\}_{i=1}^{N'}$, and all the recommendations have been obtained and checked, several metrics are used for evaluation. Accuracy (A) is computed as the fraction of successful recommendations out of the total number of words for which a recommendation was made. It is worth noting that most of the methods proposed do not provide recommendations for all the words; another interesting metric is therefore the fraction of words not covered (NC) by the system, that is, the fraction of words for which no recommendation is made. The combination of these two metrics helps us to understand how each method perform on each test set. In addition to the accuracy and the fraction of words not covered, we also compute the precision and recall as regards keeping recommendations (P_K and R_K, respectively) and changing recommendations (P_C and R_C, respectively). The latter metrics are useful since they provide specific information about the successful keeping recommendations and change recommendations, separately, while A and NC provide information about the general performance of the recommender. The code used to perform these experiments is freely available under license GNU General Public License v3.0 (Free Software Fundation, 2007) and can be downloaded from `http://transducens.dlsi.ua.es/~mespla/resources/wkr/`.

6.2 Corpora

The corpus used in our experiments is the DGT-TM (Steinberger et al., 2012). This translation memory is a collection of documents from the *Official Journal of the European Union*[18] which is aligned at the segment level for several languages (multilingual TUs). Segment alignment in DGT-TM is expected to have a high level of quality, since part of these alignments were manually checked, or were actually generated during computer-aided translation by professional translators.

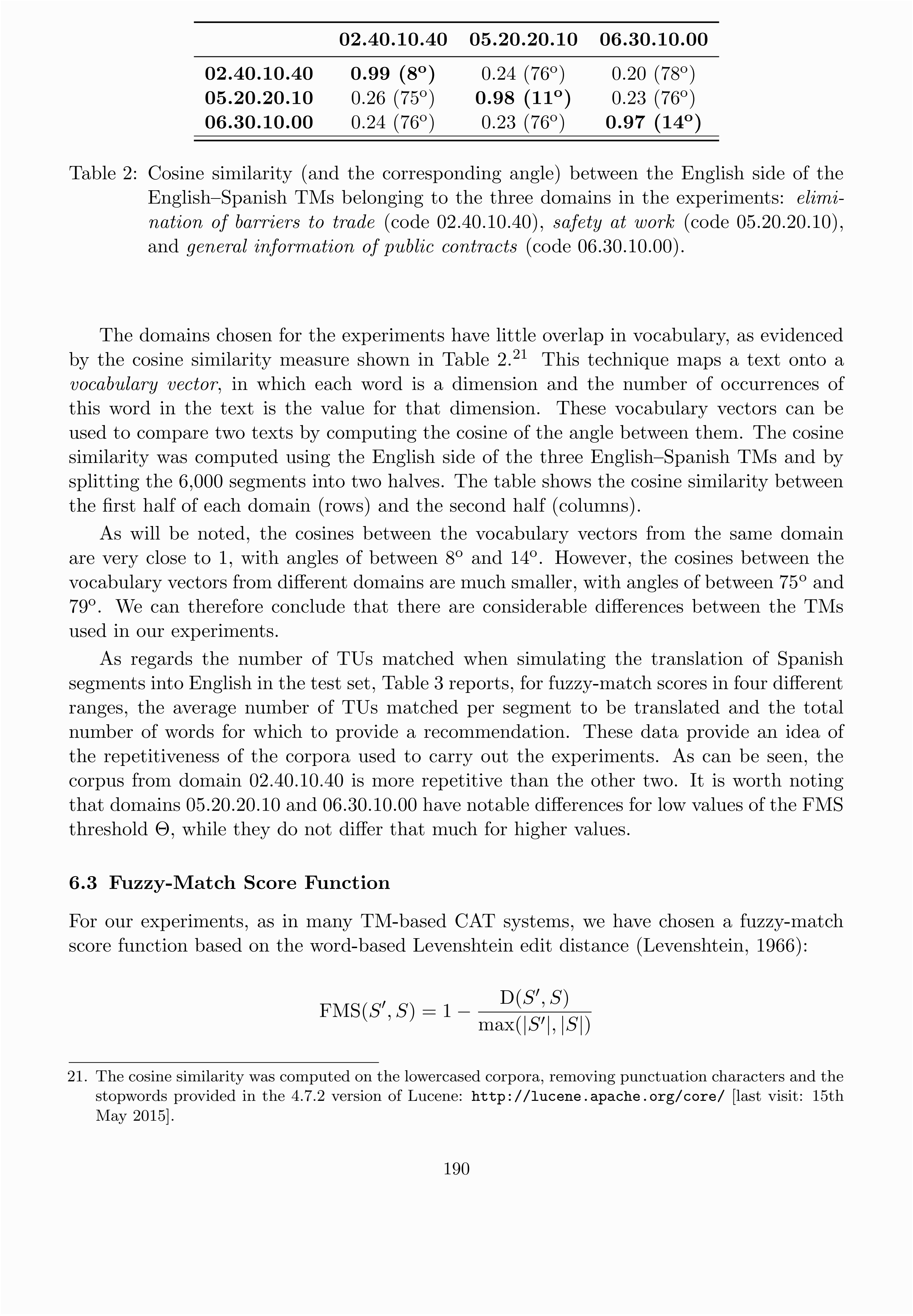
The TUs in DGT-TM contain segments in many official languages of the European Union and are labelled with domain codes[19] which were used to create three domain-specific TU collections. This was done by using the following domain codes: *elimination of barriers to trade* (code 02.40.10.40), *safety at work* (code 05.20.20.10), and *general information of public contracts* (code 06.30.10.00). Only those TUs containing the corresponding segments for all the five languages used in our experiments were included in these TU collections.

Each collection of TUs was used to build a bilingual TM and a test set for each language pair by randomly selecting pairs of segments without repetition.[20] In addition to the pre-processing already performed by the creators of DGT-TM (European Commission Joint Research Center, 2007) the segments included in the TMs and the test set used in our experiments were tokenised and lowercased. The TMs consist of $6,000$ TUs each, and simulate the TM that a translator may use when translating with a CAT tool. The test set consist of $1,500$ TUs whose source language side simulates the segments to be translated by using the TMs (the translator's job), while their target language side may be considered as a reference translation for each segment to be translated.

[18] `http://eur-lex.europa.eu` [last visit: 15th May 2015]

[19] `http://old.eur-lex.europa.eu/RECH_repertoire.do` [last visit: 15th May 2015]

[20] The TMs and the test set obtained in this way can be downloaded from `http://transducens.dlsi.ua.es/~mespla/resources/mtacat/` [last visit: 15th May 2015]



	02.40.10.40	**05.20.20.10**	**06.30.10.00**
02.40.10.40	**0.99 (8°)**	0.24 (76°)	0.20 (78°)
05.20.20.10	0.26 (75°)	**0.98 (11°)**	0.23 (76°)
06.30.10.00	0.24 (76°)	0.23 (76°)	**0.97 (14°)**

Table 2: Cosine similarity (and the corresponding angle) between the English side of the English–Spanish TMs belonging to the three domains in the experiments: *elimination of barriers to trade* (code 02.40.10.40), *safety at work* (code 05.20.20.10), and *general information of public contracts* (code 06.30.10.00).

The domains chosen for the experiments have little overlap in vocabulary, as evidenced by the cosine similarity measure shown in Table 2.[21] This technique maps a text onto a *vocabulary vector*, in which each word is a dimension and the number of occurrences of this word in the text is the value for that dimension. These vocabulary vectors can be used to compare two texts by computing the cosine of the angle between them. The cosine similarity was computed using the English side of the three English–Spanish TMs and by splitting the 6,000 segments into two halves. The table shows the cosine similarity between the first half of each domain (rows) and the second half (columns).

As will be noted, the cosines between the vocabulary vectors from the same domain are very close to 1, with angles of between 8° and 14°. However, the cosines between the vocabulary vectors from different domains are much smaller, with angles of between 75° and 79°. We can therefore conclude that there are considerable differences between the TMs used in our experiments.

As regards the number of TUs matched when simulating the translation of Spanish segments into English in the test set, Table 3 reports, for fuzzy-match scores in four different ranges, the average number of TUs matched per segment to be translated and the total number of words for which to provide a recommendation. These data provide an idea of the repetitiveness of the corpora used to carry out the experiments. As can be seen, the corpus from domain 02.40.10.40 is more repetitive than the other two. It is worth noting that domains 05.20.20.10 and 06.30.10.00 have notable differences for low values of the FMS threshold Θ, while they do not differ that much for higher values.

6.3 Fuzzy-Match Score Function

For our experiments, as in many TM-based CAT systems, we have chosen a fuzzy-match score function based on the word-based Levenshtein edit distance (Levenshtein, 1966):

$$\mathrm{FMS}(S', S) = 1 - \frac{\mathrm{D}(S', S)}{\max(|S'|, |S|)}$$

[21] The cosine similarity was computed on the lowercased corpora, removing punctuation characters and the stopwords provided in the 4.7.2 version of Lucene: `http://lucene.apache.org/core/` [last visit: 15th May 2015].

$\Theta(\%)$	**domain**	**no filtering**	
		$\mathrm{TU}_{\mathrm{avg}}$	N_{words}
≥ 60	02.40.10.40	3.71	95,881
	05.20.20.10	0.62	9,718
	06.30.10.00	5.68	34,339
≥ 70	02.40.10.40	2.36	65,865
	05.20.20.10	0.37	6,883
	06.30.10.00	0.99	10,327
≥ 80	02.40.10.40	1.58	46,519
	05.20.20.10	0.14	3,015
	06.30.10.00	0.45	4,726
≥ 90	02.40.10.40	0.70	26,625
	05.20.20.10	0.05	1,599
	06.30.10.00	0.03	1,268

Table 3: Average number of matching TUs ($\mathrm{TU}_{\mathrm{avg}}$) per segment and total number of target words (N_{words}) for which a recommendation has to be provided when translating Spanish into English for the three different domains. The results were obtained for different values of the FMS threshold (Θ).

where $|x|$ is the length (in words) of string x and $\mathrm{D}(x, y)$ refers to the word-based Levenshtein edit distance between x and y.[22]

6.4 Binary Classifier

Esplà-Gomis et al. (2011) used a simple perceptron classifier which defined, for the translation of a source segment S', the probability of keeping the j-th word in T, the target-language segment of the TU (S, T) as:

$$p_k(j, S', S, T) = \frac{1}{1 + e^{-g(j,S',S,T)}} \tag{4}$$

with

$$g(j, S', S, T) = \lambda_0 + \sum_{k=1}^{N_F} \lambda_k f_k(j, S', S, T). \tag{5}$$

This perceptron uses a sigmoid function that incorporates the linear combination of the different features f_k and the corresponding weights λ_k learned by the classifier.

[22] Many TM-based CAT tools implement variations of this FMS to rank the translation proposals as regards the edition effort required (for instance, by disregarding punctuation signs or numbers in S and S', or using stemmed versions of S and S'). In our experiments we continue to use the original FMS, since ranking is not important in our experiments. This is owing to the fact that all the proposals above the threshold are evaluated, and not only that with the highest score.

In this work, a more complex *multilayer perceptron* (Duda, Hart, & Stork, 2000, Section 6) was used, namely, that implemented in Weka 3.7 (Hall et al., 2009). Multilayer perceptrons (also known as feedforward neural networks) have a complex structure which incorporates one or more *hidden layers*, consisting of a collection H of perceptrons, placed between the input of the classifier (the features) and the output perceptron. This hidden layer makes multilayer perceptrons suitable for non-linear classification problems (Duda et al., 2000, Section 6). In fact, Hornik, Stinchcombe, and White (1989) proved that neural networks with a single hidden layer containing a finite number of neurons are universal approximators and may therefore be able to perform better than a simple perceptron for complex problems. In this case, the output perceptron that provides the classification takes the output h_l of each of the perceptrons in H as its input. Eq. (5) therefore needs to be updated as follows:

$$g(j, S', S, T) = \lambda_0 + \sum_{l=1}^{|H|} \lambda_l h_l(j, S', S, T). \tag{6}$$

Each perceptron h_l in H works similarly to the perceptron described in eq. (4):

$$h_l(j, S', S, T) = \frac{1}{1 + e^{-g_l(j,S',S,T)})}$$

with

$$g_l(j, S', S, T) = \lambda_{l0} + \sum_{k=1}^{N_F} \lambda_{lk} f_k(j, S', S, T).$$

As can be seen, besides the collection of weights $\boldsymbol{\lambda}$ for the main perceptron, a different collection of weights $\boldsymbol{\lambda}'_l$ is needed for each perceptron h_l in the hidden layer H. These weights are obtained by using the *backpropagation algorithm* (Duda et al., 2000, Section 6.3) for training, which updates them using gradient descent on the error function. In our case, we have used a batch training strategy, which iteratively updates the weights in order to minimise an error function. The training process stops when the error obtained in an iteration is worse than that obtained in the previous 10 iterations.[23]

A validation set with 10% of the training examples was used during training, and the weights were therefore iteratively updated on the basis of the error computed in the other 90%, but the decision to stop the training (usually referred as the convergence condition) was based on this validation set. This is a usual practice whose objective is to minimise the risk of overfitting.

Hyperparameter optimisation was carried out using a grid search (Bergstra, Bardenet, Bengio, & Kégl, 2011) strategy based on the accuracy obtained for the English–Spanish TM from the 02.40.10.40 domain. A 10-fold cross-validation was performed on this training corpus in order to choose the following hyperparameters:

[23] It is usual to set a number of additional iterations after the error stops improving, in case the function is in a local minimum, and the error starts decreasing again after a few more iterations. If the error continues to be worsen after these 10 iterations, the weights used are those obtained after the iteration with the lowest error.

- *Number of nodes in the hidden layer*: Weka (Hall et al., 2009) makes it possible to choose from among a collection of predefined network designs; that which best performed for our training corpus was that with the same number of nodes in the hidden layer as the number of features.
- *Learning rate*: this parameter allows the dimension of the weight updates to be regulated by applying a factor to the error function after each iteration; the value that best performed for our experiment was 0.4.
- *Momentum*: when updating the weights at the end of a training iteration, momentum modifies the new value, signifying that it not only depends on the current gradient direction, but also on the previous weight value. The objective of this technique is to smooth the training process for faster convergence. In the case of our experiments, it was set to 0.1.

6.5 Reference Results

As mentioned previously, the performance of the two MT-based approaches proposed in this work is compared to that of two different approaches: a *naïve FMS-only baseline*, which uses the classifier described in Section 6.4 and employs only the FMS between S' and S as a feature, and the approach reviewed in Section 3, which uses statistical word alignment to relate the words in the two segments of a TU (S, T). The naïve FMS-only baseline was trained on the datasets described in Section 6.2 for different values of the FMS threshold Θ. It is worth mentioning that the resulting models classify all the target words as having to be kept. This is a consequence of the fact that, for any value of the FMS in the training set, there are more words to be kept than to be changed.

The alignments used by the alignment-based approach were obtained by means of the free/open-source MGIZA++ toolkit (Gao & Vogel, 2008), an implementation of the GIZA++ toolkit (Och & Ney, 2003) which eases the task of training alignment models on a parallel corpus and then aligning a different one using the models learned. The word-based alignment models (Brown et al., 1993; Vogel, Ney, & Tillmann, 1996) were separately trained on the TMs defined in Section 6.2 and on the JRC-Acquis 3.0 (Steinberger et al., 2006) corpus (a large multilingual parallel corpus which includes, among others, texts from these TMs, given that it is built from the same texts as the DGT-TM).[24] The alignments we have used are the result of running MGIZA++ in both translation directions (source-to-target and target-to-source) and then symmetrising both sets of alignments by means of the usual *grow-diag-final-and* (Koehn et al., 2005) heuristic. This symmetrisation technique was found to be that which provided the best compromise between coverage and accuracy for word-keeping recommendation (Esplà et al., 2011).[25]

Table 4 shows the accuracy obtained by the naïve FMS-only baseline. The fraction of words not covered, that is, the words for which no recommendation is provided, is not included in this table since this baseline provides a recommendation for every word in the

[24] `https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis` [last visit: 15th May 2015]

[25] Symmetrisation is necessary because MGIZA++ produces alignments in which a source word can be aligned with many target words, whereas a target word is aligned with at most one source word. The use of symmetrisation allows alignments to be combined in both directions in order to obtain M to N alignments.

$\Theta(\%)$	$A(\%)$
≥ 60	82.69±.24
≥ 70	88.37±.25
≥ 80	91.65±.25
≥ 90	93.98±.29

Table 4: Accuracy $A(\%)$ obtained with the naïve FMS-only baseline when translating en→es in domain 02.40.10.40. Accuracy was obtained for different FMS thresholds Θ. The other language pairs and domains behave in the same way.

test set. This is due to the fact that the naïve FMS-only baseline does not depend on the coverage of a source of information.

In general, we can see that the accuracy obtained with the naïve FMS-only baseline is quite high. This is, in fact, a hard-to-beat naïve baseline, although these results are reasonable, since the relatively high values of the FMS threshold Θ imply that a high number of words should be kept unedited in the translation proposals.

With regard to the alignment-based approach, several options were evaluated in order to choose its configuration. On the one hand, we tried the two decision criteria described in Section 3 (*unanimity* and *majority*). On the other hand, we tried two alignment models: one trained on the same translation memory used for the experiments (as had occurred with the trained MT-based recommender), and another trained on the JRC Acquis parallel corpus. The objective of comparing both models was to confirm which corpus was the most adequate as regards training our alignment-based recommender: one which was reduced and domain focused, or one which was bigger and more generic, although still containing text in the same domain. The results are presented in Table 5, in which the accuracy and percentage of words not covered are measured for the four combinations of decision criteria and training corpora. As already mentioned in Section 3, the unanimity criterion is more focused on accuracy, while the majority criterion is more focused on coverage. In order to confirm which method was better, a statistical significance test was performed on the results obtained by using an approximate randomisation test.[26] The free/open-source tool SIGF V.2 (Padó, 2006) was used for the statistical significance testing of the results described throughout this section. The test confirmed that in both cases the alignment models trained on the TM used for testing outperform those trained on the JRC Acquis corpus. The approach trained on the TM used for testing will be used in all the experiments shown in the following section, while the decision criterion used will be that of unanimity, the reason being that we consider accuracy to be more relevant than coverage for word-keeping recommendation, since as mentioned above, we believe that it is better not to make a recommendation than to make a wrong one.

[26] Approximate randomisation compares the difference between the accuracy/coverage of two classifiers in the same test set. This method randomly interchanges the predictions of both classifiers in every instance of the test set. The difference between the accuracy/coverage in both randomised datasets is then compared to the original set. This process is iterativelly repeated to confirm whether the results as regards randomised predictions are consistently worse than the original results.

Θ(%)	method	training on 02.40.10.40		training on JRC Acquis	
		A (%)	NC (%)	A (%)	NC (%)
≥ 60	unanimity	**93.90±.16**	6.09±.15	93.14±.17	6.29±.15
	majority	92.96±.17	**4.39±.13**	93.05±.17	5.39±.14
≥ 70	unanimity	**94.32±.18**	5.90±.18	93.67±.19	6.15±.18
	majority	93.47±.19	**4.45±.16**	93.59±.19	5.42±.17
≥ 80	unanimity	**95.10±.20**	5.37±.21	94.56±.21	5.87±.21
	majority	94.49±.21	**4.31±.19**	94.52±.21	5.33±.20
≥ 90	unanimity	**95.34±.26**	4.93±.26	94.97±.27	5.48±.27
	majority	95.10±.27	**4.35±.25**	94.95±.27	5.04±.26

Table 5: Accuracy (A) and fraction of words not covered (NC) obtained with the alignment-based approach described in Section 3 for different FMS thresholds Θ, when translating Spanish into English in domain 02.40.10.40. The results show the accuracy obtained when using a model trained on the TM belonging to the 02.40.10.40 domain and on the JRC-Acquis corpus, both using the unanimity and the majority decision criteria. This behaviour is also observed for the remaining TMs used in the experiments. Statistically significant differences in the accuracy of each approach for the different values of Θ with $p \leq 0.05$ are highlighted in bold type, as also occurs for the fraction of words not covered.

7. Results and Discussion

In this section we present the results obtained by the two approaches proposed in this paper and compare their performance with the naïve FMS-only baseline and the alignment-based approach of Esplà et al. (2011). The large amount of variables to be taken into consideration (feature sets, language pairs, domains, MT systems, and sub-segment length) forced us to select the experiments to be performed. Some parameters were therefore chosen on the basis of the results obtained for the translation from Spanish into English, which is the language pair used by Esplà et al. (2011) and Esplà-Gomis et al. (2011). The domain chosen for these preliminary experiments is *elimination of barriers to trade* (02.40.10.40), which has higher matching rates (see Table 3) and is therefore that from which more data can be obtained.

7.1 Parameter Selection

We first attempted to determine the optimal sub-segment maximum length L for the experiments with the training-free recommender and with the trained recommender. Table 6 shows the fraction of words not covered depending on the value of L for both recommenders together. The fraction of words not covered is between 16% and 19% when using sub-segments of only one word, and the percentage diminishes as more context is provided for translations. As can be seen, the fraction of words not covered starts to stabilise with $L = 4$, since the difference between this and $L = 5$ is only about 0.25%.

Table 7 shows the impact of the value of L on the accuracy obtained by the training-free recommender and by the trained recommender when using the different sets of features

Θ(%)	**Fraction of words without recommendation** (%)				
	$L = 1$	$L = 2$	$L = 3$	$L = 4$	$L = 5$
≥ 60	16.42±.24	10.22±.19	7.24±.16	5.13±.14	4.90±.14
≥ 70	16.74±.28	10.66±.24	7.34±.20	5.18±.17	4.94±.17
≥ 80	17.37±.34	11.25±.29	7.65±.24	5.53±.21	5.29±.20
≥ 90	18.18±.46	11.80±.39	8.05±.33	5.86±.28	5.59±.28

Table 6: Percentage of words not covered by both MT-based approaches for the en→es language pair in domain 02.40.10.40 using a combination of all MT systems available. The fraction of words not covered was obtained for different FMS thresholds Θ when using different values of the maximum sub-segment length L.

Θ(%)	**Method**	**Accuracy** (%) **in classification**				
		$L = 1$	$L = 2$	$L = 3$	$L = 4$	$L = 5$
≥ 60	**MM-U**	93.51±.17	93.40±.17	93.58±.16	93.57±.16	93.77±.16
	PM-C	93.62±.17	94.07±.16	94.31±.15	94.18±.15	94.37±.15
	PM-C+C	93.59±.17	94.36±.15	94.57±.15	**95.14±.14**	95.41±.14
	training-free	93.63±.17	93.78±.16	93.79±.16	93.27±.16	92.90±.17
≥ 70	**MM-U**	94.79±.19	94.72±.18	94.77±.18	94.70±.18	94.82±.17
	PM-C	94.75±.19	94.89±.18	95.12±.17	94.94±.17	95.05±.17
	PM-C+C	94.81±.19	95.16±.17	95.33±.17	**95.63±.16**	95.92±.16
	training-free	94.76±.19	94.77±.18	94.77±.18	94.14±.18	93.78±.19
≥ 80	**MM-U**	96.09±.19	96.14±.19	95.92±.19	96.00±.18	96.02±.18
	PM-C	96.09±.19	96.14±.19	96.11±.18	96.01±.18	95.98±.18
	PM-C+C	96.11±.19	96.24±.18	96.34±.18	**96.39±.17**	96.58±.17
	training-free	96.05±.20	95.97±.19	95.88±.19	95.29±.20	94.98±.20
≥ 90	**MM-U**	96.84±.23	96.85±.22	96.80±.22	96.74±.22	96.75±.22
	PM-C	96.84±.23	96.82±.23	96.87±.22	96.85±.22	96.87±.22
	PM-C+C	96.84±.23	96.95±.22	96.95±.22	**96.90±.21**	97.00±.21
	training-free	96.80±.23	96.72±.23	96.70±.22	96.61±.22	96.42±.23

Table 7: Accuracy obtained by the trained MT-based recommender when using the different feature combinations described in Section 4 and by the training-free MT-based recommender for the en→es language pair. Accuracy was obtained for different FMS thresholds Θ when using different values of the maximum sub-segment length L. Statistically significant accuracy results for $L = 4$ (the value of L that will be used for the remaining experiments) with $p \leq 0.05$ are highlighted in bold type.

described in Section 4. As can be seen, the accuracy of the training-free system drops slightly as longer sub-segments are introduced. This is reasonable since the longer the sub-segments used, the higher the number of words for which a recommendation is made (see Table 6). Words which are covered only by very long sub-segments are more difficult to classify, since these sub-segments contain evidence regarding more words and are therefore less precise. It is interesting to observe that, in the case of most of the feature sets, the trained recommender does not behave in this manner, since it is able to learn how reliable the longer sub-segments are. In the case of the feature set MM-U, the accuracy is almost constant for all the values of L, which means that using longer sub-segments does not have an impact on the accuracy in this case. In any event, it is worth noting that the results obtained using the training-free recommender are quite accurate, which confirms that the sub-segment pairs discovered using MT are a good source of information for word-keeping recommendation. Moreover, these results indicate that long sub-segments are less informative than short sub-segments. In general, only small improvements in accuracy and coverage occur for values of L that are higher than 4. The remaining experiments in this section will therefore be performed with $L = 4$.

The results in Table 7, namely those in the column with $L = 4$, were also used to determine which is the best feature combination for the trained MT-based recommender. At first glance, the set of features based on matching words, namely PM-C (see Section 4.2), and PM-C+C (see Section 4.3), are those which perform best. As commented on in Section 4.2, MM-U features consider partial matching sub-segments as negative evidence, while PM-C and PM-C+C also attempt to extract the positive evidence from these sub-segments, thus using this bilingual information more efficiently. However, the results that they obtain are very close, particularly in the case of high values of Θ. A statistical significance test confirmed that PM-C+C is superior to all the other feature combinations for any value of Θ with $p \leq 0.05$. The feature set PM-C+C will therefore be used for the trained classifier approach in the remaining experiments in this section.

The accuracy obtained using all the approaches presented in both this and the previous section is lower than expected, particularly when considering the results obtained by Esplà-Gomis et al. (2011) and Esplà et al. (2011) in which both the trained MT-based approach and the alignment-based approach obtained an average accuracy that was about 5% higher than that obtained in our experiments. Our intuition leads us to believe that this drop in accuracy may be due to the fact that the data sets used in previous works might have been cleaner than those used here. To confirm this, an additional set of experiments was carried out using some additional cleaning criteria to ensure the quality of the datasets used for evaluation. The results of this study are presented in Appendix B.

7.2 General Results

The parameters chosen (maximum sub-segment length $L = 4$ for the MT-based approaches, PM-C+C feature set for the trained MT-based approach, and unanimity criterion and models trained on the TM to be used in the experiments for the alignment-based approach) were used to perform several experiments in order to check the performance of our system. Tables 8 and 9 show the results obtained by the trained MT-based recommender when translating Spanish segments into English. In Table 8, the different MT systems available

Θ(%)	Apertium		Google Translate		Power Translator	
	A (%)	NC (%)	A (%)	NC (%)	A (%)	NC (%)
≥ 60	94.61±.17	27.89±.28	**95.08±.14**	**6.22±.15**	94.52±.17	28.99±.29
≥ 70	95.40±.19	28.32±.34	95.54±.16	**6.38±.19**	95.47±.19	29.62±.35
≥ 80	96.49±.20	28.39±.41	96.34±.18	**6.52±.22**	96.46±.20	29.59±.42
≥ 90	**97.25±.23**	28.49±.54	96.99±.21	**6.31±.29**	97.01±.24	28.50±.54

Table 8: Accuracy (A) and fraction of words not covered (NC) obtained when translating es→en in domain 02.40.10.40 with the trained MT-based approach. The results were obtained with the separate use of the MT systems available for the language pair: Apertium, Google Translate, and Power Translator. For every value of Θ, those results that supersede the rest by a statistically significant margin of $p \leq 0.05$ are highlighted in bold type, both for accuracy and for the fraction of words not covered.

were used separately to obtain the recommendations in the 02.40.10.40 domain. The results confirm that, while accuracy remains stable,[27] coverage strongly depends on the MT system. This may be interpreted as follows: MT-based approaches are robust to bilingual sources of information with low coverage. The experiments confirmed that the best coverage is obtained with Google Translate, whereas Apertium and Power Translator produce similar results. However, Apertium and Power Translator produce higher precision for "change" recommendations, while all three MT systems perform similarly as regards the precision for "keep" recommendations.

Table 9 compares the performance of the alignment-based approach and the trained MT-based recommender when all three MT systems available are used at the same time for the language pair es→en. The table shows the results obtained separately for the alignment-based approach and the trained MT-based approach as regards the three domains: 02.40.10.40, 05.20.20.10 and 06.30.10.00. The results are quite similar for both approaches. The MT-based approach slightly outperforms the alignment-based approach in accuracy, while the results of the alignment-based approach are better for coverage, particularly in the case of the 06.30.10.00 domain. In any event, this leads us to believe that both approaches can obtain comparable results across domains.

Tables 10 and 11 present the results as regards the accuracy and fraction of words not covered, respectively, obtained with both the trained MT-based approach and the alignment-based approach for all language pairs. In the case of the trained MT-based approach, all the MT systems available were used (Table 1 lists the MT systems available for each language pair).

The results confirm the hypothesis that the alignment-based approach generally obtains better results as regards coverage than the MT-based approach for most language pairs, which is reasonable, given that the alignment models have been trained on the same TM to be used for translation. Accuracy is yet again the strongest point of the trained MT-

[27] Although some differences in accuracy can be observed, it is only possible to state which MT system is better with a statistical significance of $p \leq 0.05$ in the case of Θ=60% and Θ=90%.

Θ(%)	method	02.40.10.40		05.20.20.10		06.30.10.00	
		A (%)	NC (%)	*A* (%)	NC (%)	*A* (%)	NC (%)
≥ 60	trained	**95.14±.1**	**5.13±.1**	**95.62±.4**	7.20±.5	**94.87±.3**	**11.33±.3**
	alignment	93.90±.2	6.09±.2	92.97±.5	**6.22±.5**	91.38±.3	18.31±.4
≥ 70	trained	**95.63±.2**	**5.18±.2**	**97.02±.4**	6.44±.6	94.93±.5	10.55±.6
	alignment	94.32±.2	5.90±.2	94.16±.6	5.96±.6	94.57±.5	**7.78±.5**
≥ 80	trained	**96.39±.2**	5.53±.2	**95.78±.8**	10.42±1.1	95.00±.7	12.84±1
	alignment	95.10±.2	5.37±.2	94.56±.8	**5.51±.8**	95.20±.6	**4.80±.6**
≥ 90	trained	**96.90±.2**	5.86±.3	95.36±1.1	11.13±1.5	97.65±.9	9.31±1.6
	alignment	95.34±.3	**4.93±.3**	94.26±1.2	**5.19±1.1**	96.47±1.1	**6.15±1.3**

Table 9: Accuracy (A) and fraction of words not covered (NC) obtained with the trained MT-based approach and the alignment-based approach when translating es→en in the three different domains and using all the MT systems. For each corpus and for each value of Θ, a statistical significance test was performed for both approaches as regards both the accuracy and the fraction of words not covered. Those results that show an improvement which is statistically significant with $p \leq 0.05$ are highlighted in bold type.

	Θ ≥ 60(%)		Θ ≥ 70(%)		Θ ≥ 80(%)		Θ ≥ 90(%)	
lang. pair	trained	alignment	trained	alignment	trained	alignment	trained	alignment
es→en	**95.1±.1**	93.9±.2	**95.5±.2**	94.3±.2	**96.3±.2**	95.1±.2	**97.0±.2**	95.3±.3
en→es	**90.0±.2**	88.7±.2	**91.5±.2**	89.8±.2	**92.3±.2**	90.4±.2	**93.0±.2**	91.8±.3
de→en	**94.2±.2**	92.6±.2	**95.8±.2**	93.8±.2	**96.6±.2**	94.7±.2	**97.2±.2**	95.5±.3
en→de	**88.9±.2**	87.9±.2	**91.1±.2**	89.7±.2	**93.0±.2**	91.5±.2	**92.8±.3**	91.5±.3
fr→en	**95.2±.1**	93.4±.2	**96.3±.2**	95.1±.2	**97.0±.2**	96.0±.2	**97.6±.2**	96.4±.2
en→fr	**89.7±.2**	89.4±.2	**91.9±.2**	91.3±.2	**93.3±.2**	92.1±.2	**95.7±.2**	94.3±.3
fi→en	**93.2±.2**	91.9±.2	**94.6±.2**	93.1±.2	**94.7±.2**	93.3±.3	**94.8±.3**	92.8±.4
en→fi	**89.1±.2**	87.1±.2	**90.2±.3**	88.5±.3	**90.3±.3**	88.4±.3	**90.6±.4**	88.9±.4
es→fr	**91.2±.2**	89.4±.2	**93.2±.2**	90.7±.2	**94.8±.2**	92.8±.2	**96.0±.2**	94.3±.2
fr→es	**89.4±.2**	88.6±.2	**92.1±.2**	91.1±.2	**93.5±.2**	92.2±.2	**93.6±.3**	92.2±.3

Table 10: Accuracy (A) obtained with both the trained MT-based approach and the alignment-based approach when translating between all the language pairs in domain 02.40.10.40. The results were obtained for several values of the FMS threshold Θ and using all available MT systems for each language pair. For each language pair and for each value of Θ, a statistical significance test was performed between the accuracy obtained by both approaches. Those results that show an improvement which is statistically significant with $p \leq 0.05$ are highlighted in bold type.

	$\Theta \geq 60(\%)$		$\Theta \geq 70(\%)$		$\Theta \geq 80(\%)$		$\Theta \geq 90(\%)$	
lang. pair	trained	align-ment	trained	align-ment	trained	align-ment	trained	align-ment
es→en	6.2±.1	6.1±.1	6.4±.2	**5.9±.2**	6.5±.2	**5.4±.2**	6.3±.3	**4.9±.3**
en→es	**7.3±.1**	7.7±.1	8.3±.2	8.8±.2	**8.6±.2**	9.4±.2	**9.5±.3**	10.7±.3
de→en	10.5±.2	**5.8±.1**	10.7±.2	**6.2±.2**	10.9±.3	**6.0±.2**	12.4±.4	**6.0±.3**
en→de	11.6±.2	**5.6±.1**	12.4±.2	**6.2±.2**	13.0±.3	**6.5±.2**	14.5±.4	**6.6±.3**
fr→en	7.7±.2	**4.3±.1**	7.7±.2	**4.1±.2**	8.1±.3	**4.2±.2**	7.6±.3	**3.5±.2**
en→fr	7.9±.1	**5.9±.1**	8.8±.2	**6.0±.2**	9.6±.2	**6.0±.2**	9.4±.3	**7.8±.3**
fi→en	11.2±.2	**9.9±.2**	11.3±.3	**10.2±.2**	11.5±.3	**10.6±.3**	11.6±.4	**10.9±.4**
en→fi	11.6±.2	**7.2±.2**	11.0±.3	**7.0±.2**	12.0±.3	**7.2±.2**	12.6±.4	**8.0±.3**
es→fr	17.8±.2	**4.9±.1**	18.4±.2	**4.9±.1**	18.6±.3	**5.0±.2**	21.2±.4	**5.6±.2**
fr→es	19.5±.2	**10.3±.1**	19.4±.3	**9.5±.2**	20.4±.3	**10.5±.2**	17.6±.4	**5.9±.2**

Table 11: Fraction of words not covered (NC) obtained with both the trained MT-based approach and the alignment-based approach when translating between all the language pairs in domain 02.40.10.40. The results were obtained for several values of the FMS threshold Θ and using all available MT systems for each language pair. For each language pair and for each value of Θ, a statistical significance test was performed between the fraction of words not covered obtained by both approaches. Those results that show an improvement which is statistically significant with $p \leq 0.05$ are highlighted in bold type.

based recommender. Another interesting detail is that the experiments for language pairs with English as the target language obtain better accuracy than the experiments for the inverse language pairs. This is due to the fact that in the DGT-TM, it is usual to find *free translations* in languages other than English, which is the original language of most of the documents in this TM.[28] It is particularly frequent to find additional information about technical English words in the other languages. For example, when *software* is translated into Spanish, the translated segments include the text *"equipo lógico" ("software").* The translation includes both the correct translation and the original word, in order to keep the meaning of the original English segment. This is an important issue since, when these free translations are used as a reference to evaluate the accuracy of the approaches presented in this work, they lead to lower accuracy. This problem is analysed, and partially bypassed, in the additional experiments presented in Appendix B. As can be seen, the trained MT-based approaches provide, in general, better accuracy. However, for most of the language pairs, the coverage obtained with the alignment-based approach is much better. Nevertheless the results are still reasonably similar and can be considered comparable.

Table 12 shows the results obtained with the different approaches:[29] the naïve FMS-only baseline, the alignment-based approach, and both the trained and the training-free MT-based approaches, for the es→en language pair in the 02.40.10.40 domain and using all the MT systems available simultaneously. This table provides more detail: in addition to the accuracy and percentage of words not covered, it also includes the precision and recall for both keeping recommendations and changing recommendations. This table allows to better understand the differences between the approaches before starting a more complex comparison in different scenarios. Throughout this section we provide information regarding precision and recall where it is significant for all the tables presented.

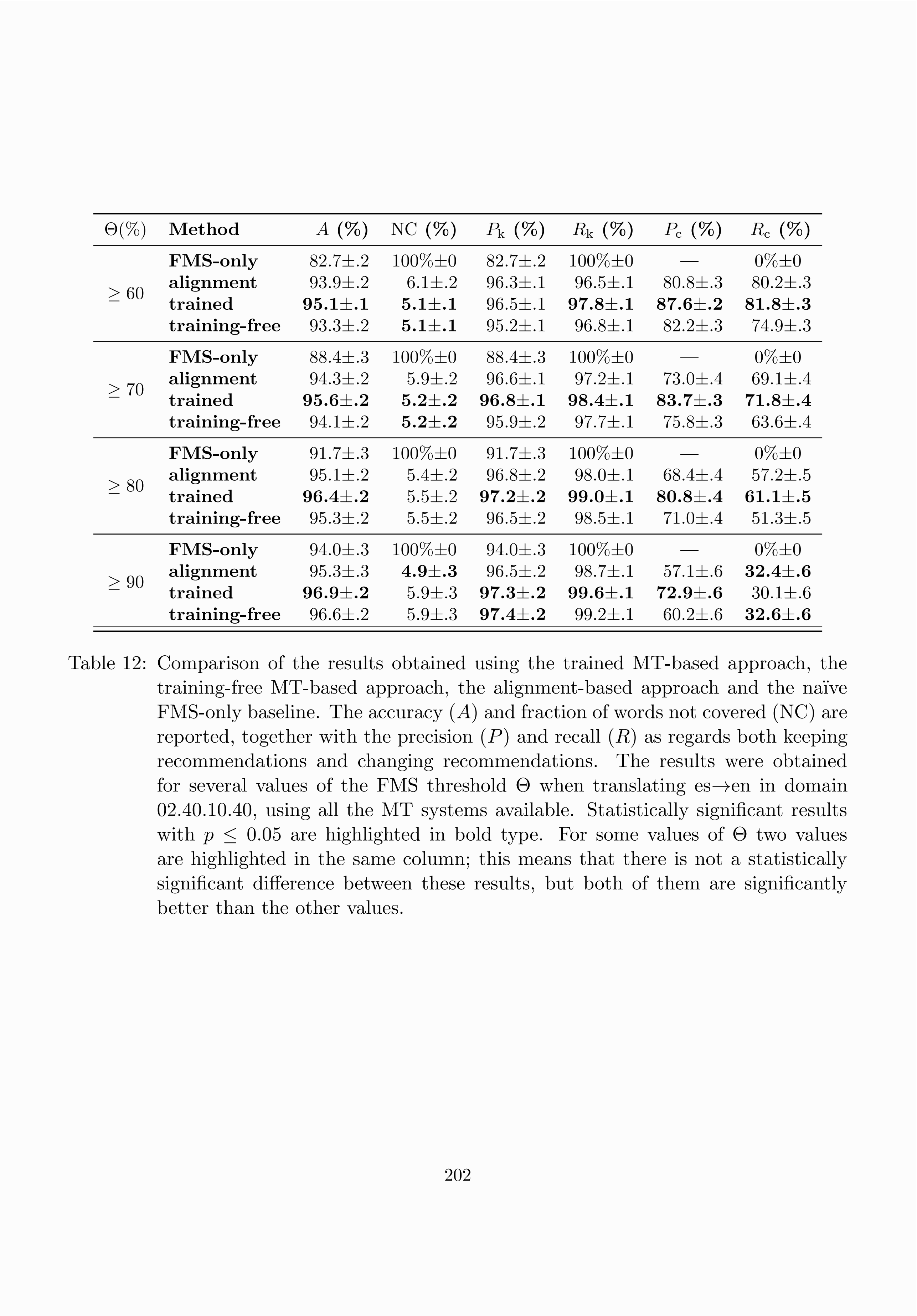
Leaving aside the naïve FMS-only baseline, it can be observed that the accuracy is similar for all the approaches, and that the trained MT-based recommender obtains slightly better results. As already mentioned, the amount of words not covered is very similar for the MT-based approaches and the alignment-based approach. As regards precision, the trained MT-based approach seems to outperform the others, although all the approaches obtain comparable scores. These results are coherent with those obtained for the rest of language pairs: in general recall and precision in "keep" recommendations are similar for both approaches, while the MT-based approach seems to be more precise in the case of "change" recommendations, where the difference is much higher, specially for higher values of the FMS threshold Θ. These conclusions are extensible to the data shown in Table 13.

The results shown in Table 12 were extended by repeating the experiment and computing recommendations only for content words (i.e. ignoring stopwords). This was done by using the list of stopwords provided in the 4.7.2 version of Lucene[30] for the language utilised in our experiments. The results of this experiments can be found in Table 13. As can be seen, the results do not change much, thus confirming that all the approaches perform equally well with content and stopwords.

[28] According to Steinberger et al. (2012), English is the source language in 72% of the documents.

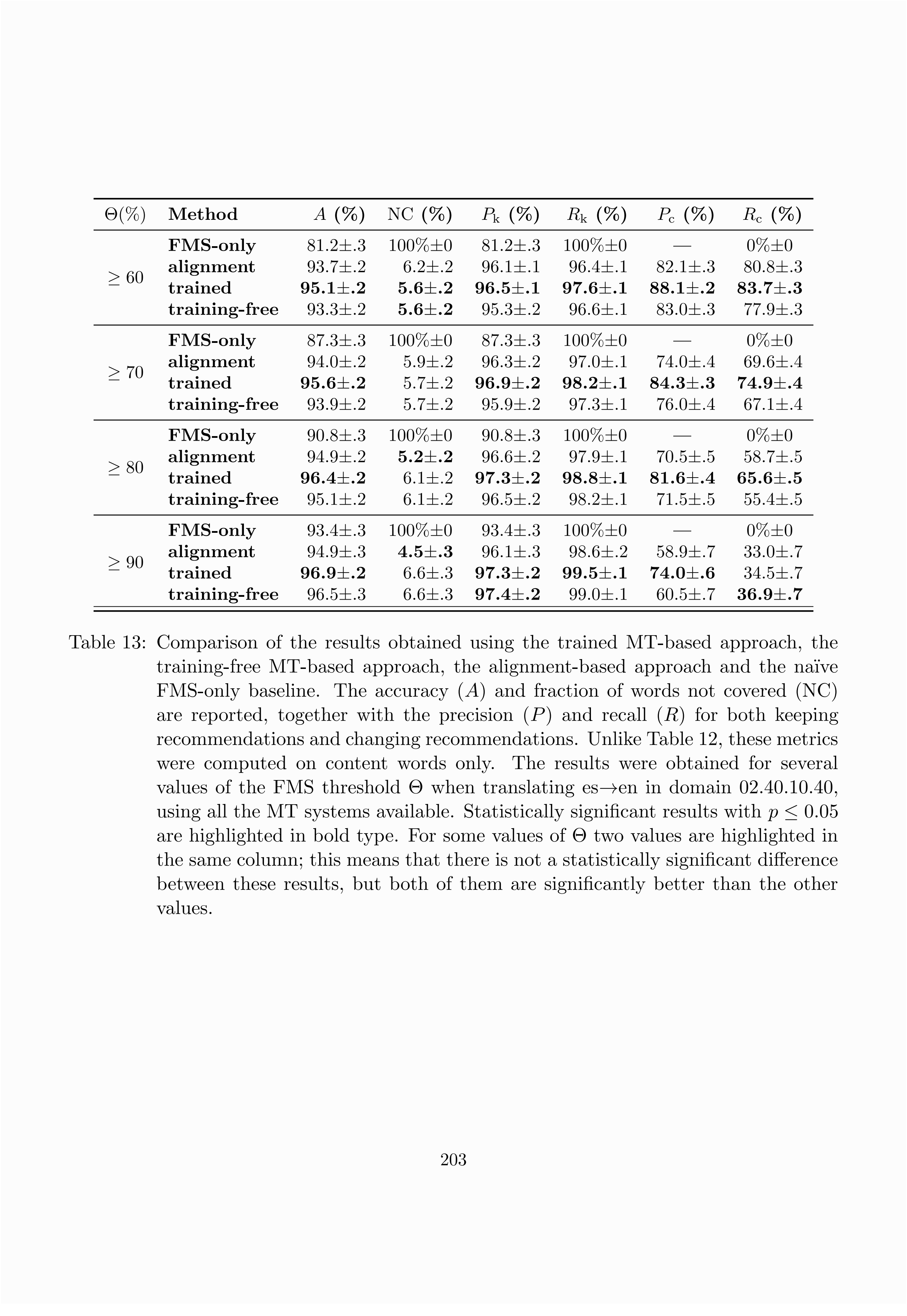
[29] The naïve FMS-only baseline did not recommend that any word be changed, as explained in Section 6.5, signifying that P_c cannot be computed for this approach and R_c is always 0. Similarly, all the words in this approach are covered, and therefore R_k is always 1.

[30] `http://lucene.apache.org/core/` [last visit: 15th May 2015]



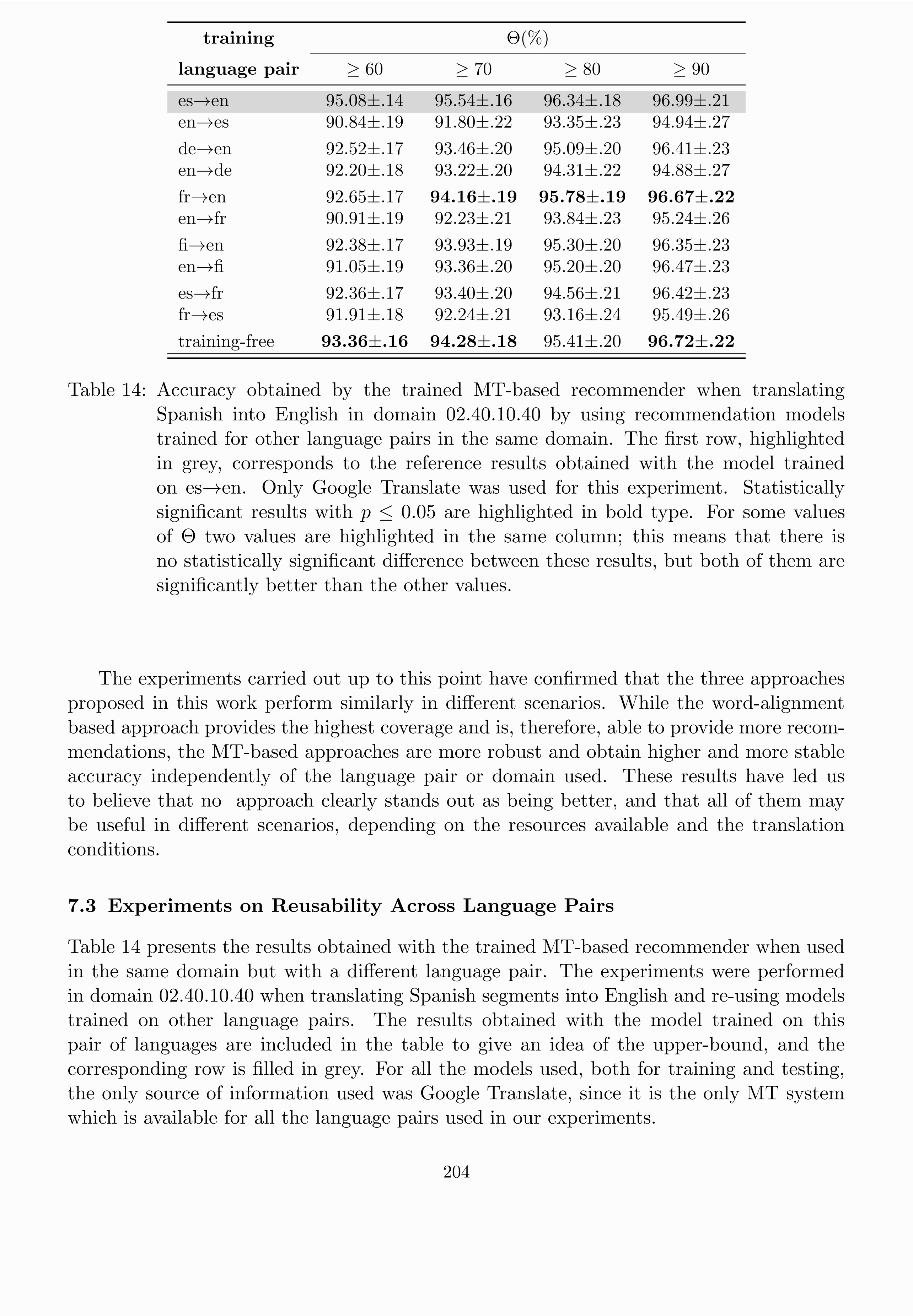
Θ(%)	Method	A (%)	NC (%)	P_k (%)	R_k (%)	P_c (%)	R_c (%)
≥ 60	**FMS-only**	82.7±.2	100%±0	82.7±.2	100%±0	—	0%±0
	alignment	93.9±.2	6.1±.2	96.3±.1	96.5±.1	80.8±.3	80.2±.3
	trained	**95.1±.1**	**5.1±.1**	96.5±.1	**97.8±.1**	**87.6±.2**	**81.8±.3**
	training-free	93.3±.2	**5.1±.1**	95.2±.1	96.8±.1	82.2±.3	74.9±.3
≥ 70	**FMS-only**	88.4±.3	100%±0	88.4±.3	100%±0	—	0%±0
	alignment	94.3±.2	5.9±.2	96.6±.1	97.2±.1	73.0±.4	69.1±.4
	trained	**95.6±.2**	**5.2±.2**	**96.8±.1**	**98.4±.1**	**83.7±.3**	**71.8±.4**
	training-free	94.1±.2	**5.2±.2**	95.9±.2	97.7±.1	75.8±.3	63.6±.4
≥ 80	**FMS-only**	91.7±.3	100%±0	91.7±.3	100%±0	—	0%±0
	alignment	95.1±.2	5.4±.2	96.8±.2	98.0±.1	68.4±.4	57.2±.5
	trained	**96.4±.2**	5.5±.2	**97.2±.2**	**99.0±.1**	**80.8±.4**	**61.1±.5**
	training-free	95.3±.2	5.5±.2	96.5±.2	98.5±.1	71.0±.4	51.3±.5
≥ 90	**FMS-only**	94.0±.3	100%±0	94.0±.3	100%±0	—	0%±0
	alignment	95.3±.3	**4.9±.3**	96.5±.2	98.7±.1	57.1±.6	**32.4±.6**
	trained	**96.9±.2**	5.9±.3	**97.3±.2**	**99.6±.1**	**72.9±.6**	30.1±.6
	training-free	96.6±.2	5.9±.3	**97.4±.2**	99.2±.1	60.2±.6	**32.6±.6**

Table 12: Comparison of the results obtained using the trained MT-based approach, the training-free MT-based approach, the alignment-based approach and the naïve FMS-only baseline. The accuracy (A) and fraction of words not covered (NC) are reported, together with the precision (P) and recall (R) as regards both keeping recommendations and changing recommendations. The results were obtained for several values of the FMS threshold Θ when translating es→en in domain 02.40.10.40, using all the MT systems available. Statistically significant results with $p \leq 0.05$ are highlighted in bold type. For some values of Θ two values are highlighted in the same column; this means that there is not a statistically significant difference between these results, but both of them are significantly better than the other values.



Θ(%)	**Method**	A (%)	NC (%)	P_{k} (%)	R_{k} (%)	P_{c} (%)	R_{c} (%)
≥ 60	**FMS-only**	81.2±.3	100%±0	81.2±.3	100%±0	—	0%±0
	alignment	93.7±.2	6.2±.2	96.1±.1	96.4±.1	82.1±.3	80.8±.3
	trained	**95.1±.2**	**5.6±.2**	**96.5±.1**	**97.6±.1**	**88.1±.2**	**83.7±.3**
	training-free	93.3±.2	**5.6±.2**	95.3±.2	96.6±.1	83.0±.3	77.9±.3
≥ 70	**FMS-only**	87.3±.3	100%±0	87.3±.3	100%±0	—	0%±0
	alignment	94.0±.2	5.9±.2	96.3±.2	97.0±.1	74.0±.4	69.6±.4
	trained	**95.6±.2**	5.7±.2	**96.9±.2**	**98.2±.1**	**84.3±.3**	**74.9±.4**
	training-free	93.9±.2	5.7±.2	95.9±.2	97.3±.1	76.0±.4	67.1±.4
≥ 80	**FMS-only**	90.8±.3	100%±0	90.8±.3	100%±0	—	0%±0
	alignment	94.9±.2	**5.2±.2**	96.6±.2	97.9±.1	70.5±.5	58.7±.5
	trained	**96.4±.2**	6.1±.2	**97.3±.2**	**98.8±.1**	**81.6±.4**	**65.6±.5**
	training-free	95.1±.2	6.1±.2	96.5±.2	98.2±.1	71.5±.5	55.4±.5
≥ 90	**FMS-only**	93.4±.3	100%±0	93.4±.3	100%±0	—	0%±0
	alignment	94.9±.3	**4.5±.3**	96.1±.3	98.6±.2	58.9±.7	33.0±.7
	trained	**96.9±.2**	6.6±.3	**97.3±.2**	**99.5±.1**	**74.0±.6**	34.5±.7
	training-free	96.5±.3	6.6±.3	**97.4±.2**	99.0±.1	60.5±.7	**36.9±.7**

Table 13: Comparison of the results obtained using the trained MT-based approach, the training-free MT-based approach, the alignment-based approach and the naïve FMS-only baseline. The accuracy (A) and fraction of words not covered (NC) are reported, together with the precision (P) and recall (R) for both keeping recommendations and changing recommendations. Unlike Table 12, these metrics were computed on content words only. The results were obtained for several values of the FMS threshold Θ when translating es→en in domain 02.40.10.40, using all the MT systems available. Statistically significant results with $p \leq 0.05$ are highlighted in bold type. For some values of Θ two values are highlighted in the same column; this means that there is not a statistically significant difference between these results, but both of them are significantly better than the other values.



training	Θ(%)			
language pair	≥ 60	≥ 70	≥ 80	≥ 90
es→en	95.08±.14	95.54±.16	96.34±.18	96.99±.21
en→es	90.84±.19	91.80±.22	93.35±.23	94.94±.27
de→en	92.52±.17	93.46±.20	95.09±.20	96.41±.23
en→de	92.20±.18	93.22±.20	94.31±.22	94.88±.27
fr→en	92.65±.17	**94.16±.19**	**95.78±.19**	**96.67±.22**
en→fr	90.91±.19	92.23±.21	93.84±.23	95.24±.26
fi→en	92.38±.17	93.93±.19	95.30±.20	96.35±.23
en→fi	91.05±.19	93.36±.20	95.20±.20	96.47±.23
es→fr	92.36±.17	93.40±.20	94.56±.21	96.42±.23
fr→es	91.91±.18	92.24±.21	93.16±.24	95.49±.26
training-free	**93.36±.16**	**94.28±.18**	95.41±.20	**96.72±.22**

Table 14: Accuracy obtained by the trained MT-based recommender when translating Spanish into English in domain 02.40.10.40 by using recommendation models trained for other language pairs in the same domain. The first row, highlighted in grey, corresponds to the reference results obtained with the model trained on es→en. Only Google Translate was used for this experiment. Statistically significant results with $p \leq 0.05$ are highlighted in bold type. For some values of Θ two values are highlighted in the same column; this means that there is no statistically significant difference between these results, but both of them are significantly better than the other values.

The experiments carried out up to this point have confirmed that the three approaches proposed in this work perform similarly in different scenarios. While the word-alignment based approach provides the highest coverage and is, therefore, able to provide more recommendations, the MT-based approaches are more robust and obtain higher and more stable accuracy independently of the language pair or domain used. These results have led us to believe that no approach clearly stands out as being better, and that all of them may be useful in different scenarios, depending on the resources available and the translation conditions.

7.3 Experiments on Reusability Across Language Pairs

Table 14 presents the results obtained with the trained MT-based recommender when used in the same domain but with a different language pair. The experiments were performed in domain 02.40.10.40 when translating Spanish segments into English and re-using models trained on other language pairs. The results obtained with the model trained on this pair of languages are included in the table to give an idea of the upper-bound, and the corresponding row is filled in grey. For all the models used, both for training and testing, the only source of information used was Google Translate, since it is the only MT system which is available for all the language pairs used in our experiments.

The results show a clear decline as regards the results obtained when the recommendation model is learned for es→en and when it is learned from the other language pairs, particularly for low values of the FMS threshold Θ. In most cases, the accuracy obtained when training the recommender on language pairs that are different from those used for testing is worse than that obtained using the training-free approach. The only exception is the model trained on the fr→en pair, which is the most similar pair to es→en. The statistical significance test confirms that, for all the values of Θ, either this model or the training-free approach are the best ones. The difference between the accuracy obtained for these approaches for Θ=70% and for Θ=90% is not in fact statistically significant.

These results have led us to believe that the trained method is highly dependent on the language pair used for training, thus making it reasonable to conclude that it is better to use the training-free MT-based recommender than an MT-based recommender trained on a different language pair.

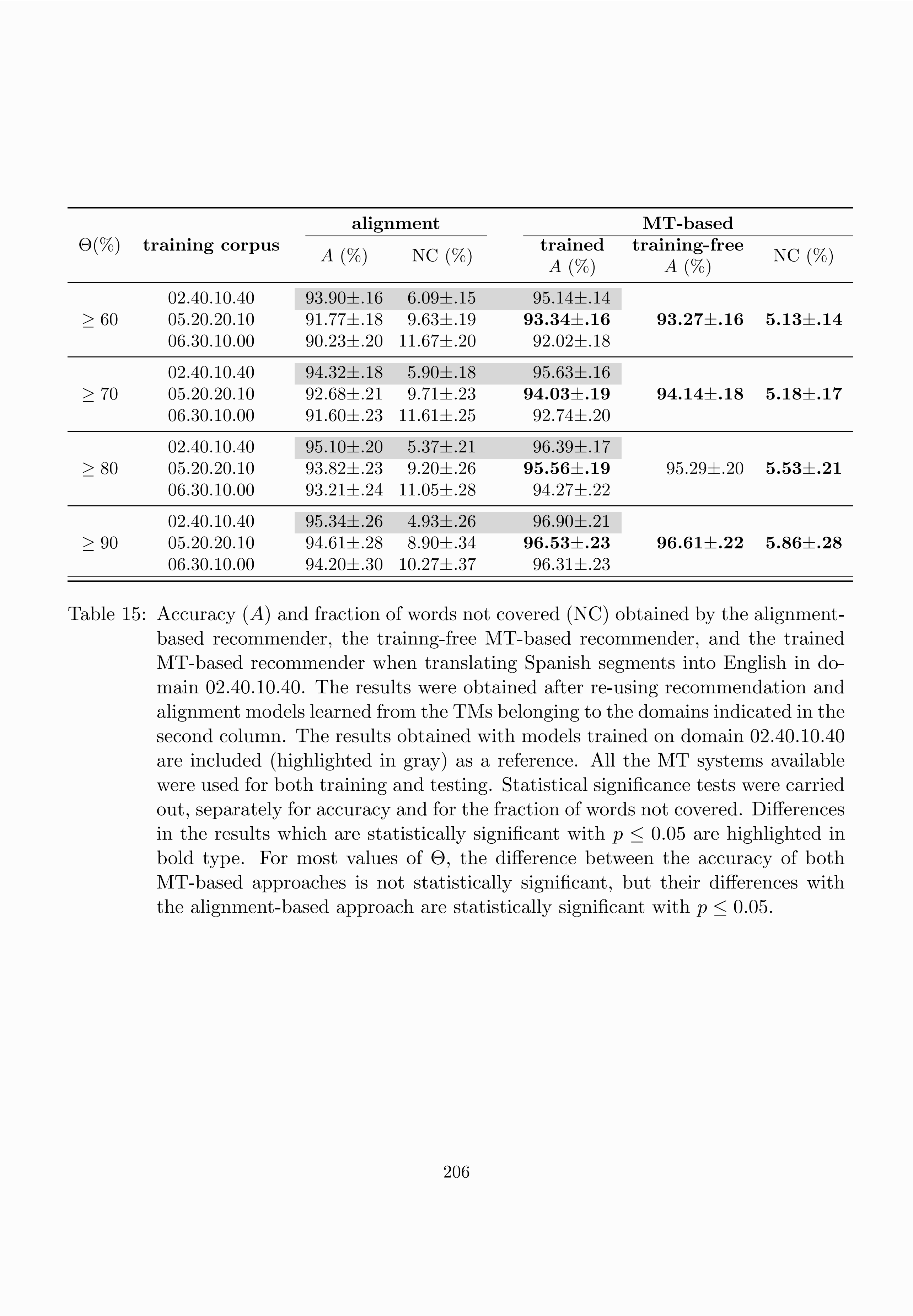
7.4 Experiments on Reusability Across Domains

Table 15 presents the results of the experiments concerning domain independence. The objective of these experiments is to verify how dependent the trained MT-based recommender is on the domain of the training corpus. In this case, we re-used the recommendation models trained in the three domains for the es→en translation to translate Spanish segments from domain 02.40.10.40 into English, and using all the MT systems available.

A drop in accuracy can be observed when re-using models trained on out-of-domain TMs rather than training on the TM to be used for translation. However, in this case the accuracy is closer to that obtained when the recommendation model is trained on the same TM used for testing (in-domain). With regard to the results obtained with the alignment-based approach, the difference in accuracy of all the MT-based approaches is higher and statistically significant with $p \leq 0.05$. That is to say, training-free MT-based approach and the models trained on domain 05.20.20.10 are those which perform best, with no statistically significant difference for most values of Θ. Similarly, the coverage of the alignment-based approach clearly drops when using out-of-domain models. This is due to the fact that, in the case of the alignment-based approach, those words which were not seen during training cannot be aligned, since no translation probabilities are learned for them. In contrast, in the case of the MT-based approaches the linguistic resources are not learned during training, but are rather obtained from the MT systems available: the training of the MT-based recommender instead focuses on the relevance of sub-segment lengths and the amount of sub-segment pairs covering each word. In general, the conclusion drawn from this experiment is that using either the training-free approach or a classification model trained on a corpus from a different domain are both valid options and they perform better than the alignment-based approach. Having a closer look to the data one can observe that the bigger differences are in the precision for "change" recommendations, where the MT-based approach outperforms the alignment-based approach.

7.5 Experiments on Reusability Across Machine Translation Systems

Table 16 presents the results of the experiments concerning MT system independence. Three models were trained on the three TM belonging to domain 02.40.10.40, but in each case



Θ(%)	training corpus	alignment		MT-based		
		A (%)	NC (%)	trained A (%)	training-free A (%)	NC (%)
≥ 60	02.40.10.40	93.90±.16	6.09±.15	95.14±.14	**93.27±.16**	**5.13±.14**
	05.20.20.10	91.77±.18	9.63±.19	**93.34±.16**		
	06.30.10.00	90.23±.20	11.67±.20	92.02±.18		
≥ 70	02.40.10.40	94.32±.18	5.90±.18	95.63±.16	**94.14±.18**	**5.18±.17**
	05.20.20.10	92.68±.21	9.71±.23	**94.03±.19**		
	06.30.10.00	91.60±.23	11.61±.25	92.74±.20		
≥ 80	02.40.10.40	95.10±.20	5.37±.21	96.39±.17	95.29±.20	**5.53±.21**
	05.20.20.10	93.82±.23	9.20±.26	**95.56±.19**		
	06.30.10.00	93.21±.24	11.05±.28	94.27±.22		
≥ 90	02.40.10.40	95.34±.26	4.93±.26	96.90±.21	**96.61±.22**	**5.86±.28**
	05.20.20.10	94.61±.28	8.90±.34	**96.53±.23**		
	06.30.10.00	94.20±.30	10.27±.37	96.31±.23		

Table 15: Accuracy (A) and fraction of words not covered (NC) obtained by the alignment-based recommender, the trainng-free MT-based recommender, and the trained MT-based recommender when translating Spanish segments into English in domain 02.40.10.40. The results were obtained after re-using recommendation and alignment models learned from the TMs belonging to the domains indicated in the second column. The results obtained with models trained on domain 02.40.10.40 are included (highlighted in gray) as a reference. All the MT systems available were used for both training and testing. Statistical significance tests were carried out, separately for accuracy and for the fraction of words not covered. Differences in the results which are statistically significant with $p \leq 0.05$ are highlighted in bold type. For most values of Θ, the difference between the accuracy of both MT-based approaches is not statistically significant, but their differences with the alignment-based approach are statistically significant with $p \leq 0.05$.

Θ(%)	A (%) trained			A (%) training-free
	Apertium	**Google Translate**	**Power Translator**	
≥ 60	92.94±.17	95.08±.14	91.12±.19	**93.36±.16**
≥ 70	93.41±.20	95.54±.16	93.02±.20	**94.28±.18**
≥ 80	**95.57±.19**	96.34±.18	94.78±.21	95.41±.20
≥ 90	96.65±.22	96.99±.21	96.47±.23	96.72±.22

Table 16: Accuracy (A) obtained by the trained MT-based recommender and the training-free MT-based recommender when translating Spanish segments into English in domain 02.40.10.40 and using Google Translate as the MT system for testing. For the trained approach, the results were obtained after re-using recommendation models learned from the same TM and language pairs but using the three different MT systems. The results obtained using a model trained on Google (column in gray) are included as an upper-bound, but are not included in the comparison. For every value of Θ the highest accuracy, with a statistically significant difference with $p \leq 0.05$ compared to the other values, is highlighted in bold type.

using one of the three MT systems available. The models were used to translate segments in Spanish into English within the same domain and using Google Translate as the MT system, in order to obtain the sub-segment translations during testing. The results in this table are similar to those presented in the last set of experiments, in which the reusability across different domains was studied. In general terms, it would appear that the drop in accuracy when making "change" recommendations is quite similar for the models trained on Apertium and Power Translator. In addition, we observed that the accuracy obtained for these two models is similar to that obtained by the training-free recommender. The training-free approach is, in fact, that which performs best for $\Theta \geq 60\%$ and $\Theta \geq 70\%$ and the difference in accuracy is statistically significant with $p \leq 0.05$. However, for $\Theta \geq 80\%$ the trained approach using a model trained with Apertium is that which performs best. Finally, there is no difference among the results for $\Theta = 90\%$.

In general, it would appear that re-using models trained on an MT system which is different to that used for translation is feasible, although using the training-free approach can provide better results.

7.6 Error Analysis

The following is a sample of the most frequent errors made by the different approaches proposed in this work for word-keeping recommendation. The objective of this error analysis is to propose strategies to deal with these errors (when possible) in future work.

7.6.1 Errors Caused by Synonyms or Equivalent Expressions

Some of the incorrect change recommendations in our experiments resulted from the use of different synonyms in the translation proposal T and the reference T' used as a gold standard. Let us suppose a translation proposal (S, T)= (*"the natural isotopic abundance* *of lithium-6 is approximately 6,5 weight per cent (7,5 atom per cent).", "la proporción natural del isótopo litio-6 es de aproximadamente 6,5% del peso (7,5% de átomos).")*, for the sentence S'= *"the natural isotopic abundance of lithium-6 is approximately 6,5 weight % (7,5 atom %)."* whose reference translation is T'= *"la proporción natural del isótopo 6 en el litio es de aproximadamente 6,5% en peso (7,5% de átomos).".* As can be seen, S and S' are semantically equivalent and are written almost the same, although the percentage symbol (%) is used in S', while the expression *per cent* is used in S. Although these two options are equivalent, *per cent* is not considered to be part of the matching between S and S' in any of the occurrences. A sub-segment pair (σ, τ) with σ = *"per cent"* and τ = *"%"*, may have led the two occurrences of the symbol % in T to be changed, when this is obviously not necessary. Since the symbol % was also used in the reference translation T', these are in fact considered to be wrong recommendations in the evaluation.

7.6.2 Errors Caused by Morphological Differences Between the Languages

This problem is in some respects similar to the previous one, although here the problem does not concern using different words for the same concept, but rather the presence of a word in one of the languages that may have different morphologies in the other language. For example, we found the proposal (S, T)= *("optical equipment as follows:","equipo óptico según se indica:")*, for the sentence S'= *"optical detectors, as follows:"* whose reference translation is T'= *"detectores ópticos según se indica:".* In this case the word *optical* is matched in both S and S', being singular in S and plural in S'. While in English both forms share the same orthography, in Spanish, the plural mark is added in T' (*ópticos*), therefore differing from the singular form in T (*óptico*). As a result, the word *óptico* would be probably recommended to be "kept", although it is not actually in the final translation (or at least not inflected in this way). It is worth noting that this would not be such a bad recommendation, since the difference between the word to be kept and the word needed for the translation is the same, but inflected in a different way. Whatever the case might be, it would be necessary to indicate this situations in some way, in order to let the user know that a change must be made.

7.6.3 Errors Caused by Fertility

This refers to the fact that the translation of a single word in one language is translated by two or more word in the other language. These words form a multi-word expression that can be translated properly only when using a sub-segment covering the whole expression. Sub-segments covering only a part of the expression can lead to out-of-context translations that produce wrong evidence. For instance, in the TU (S, T)= *("wavelength of less than 1400nm", "longitud de onda inferior a 1400nm")*, proposed for the sentence S'= *"wavelength equal to or greater than 1400nm"* whose reference translation is T'= *"longitud de onda igual o superior a 1400nm"* the word *wavelength* in English is translated as *longitud de onda* in Spanish. Since *wavelength* appears in both S and S', it is obvious that the three words in this multi-word expression should be kept. However, out of this context, word *de* can be translated as *of*, which also appears in S and is not matched in S'. The word *de* may therefore be obtaining "keeping evidence" from the sub-segments covering the whole expression *longitud de onda*, but "changing evidence" from the sub-segments covering only a part of the expression. This situation may easily result in a change recommendation. This is probably the most difficult error to fix, since it is motivated by the specifics of each language and may, in some cases, be extremely complex.

8. Concluding Remarks

In this paper we have presented a new approach to assist CAT users who employ TMs by providing them with recommendations as to which target-side words in the translation proposals provided by the CAT system have to be changed (modified or removed) or kept unedited. The method we propose imposes no constraints on the type of MT system to be used, since it is used as a black box. This method may use more than one MT system at the same time to obtain a set of features that are then combined using a binary classifier to determine whether target words have to be changed or should remain unedited. In any event, MT results are never presented to the translator. A version of this method which does not require any training is also proposed as an alternative.

The experiments carried out bear witness to the feasibility of the methods proposed by comparing the accuracy and coverage to those of a previous approach based on statistical word alignment for word-keeping recommendation. These experiments tackle the problem in different scenarios, comparing the results obtained for different domains, MT systems and language pairs. The results obtained confirm the viability of the MT-based methods proposed in this work, particularly in the case of the trained MT-based approach (see Section 4), which obtains better results as regards accuracy than those obtained with the statistical alignment-based approach (see Section 3). In the case of coverage, the results obtained with the MT-based approaches are in general worse than those obtained using the alignment-based approach when the alignment models are trained on in-domain TMs, but better when they are trained on out-of-domain TMs. The results also show a reasonable degree of independence of the MT-based models with respect to the domain of the TM and the MT system(s) used for training. These results suggest that there is no need to re-train the classifier for each new TM and, even more importantly, that it is not necessary to do so every time a new TU is added to the TM.

In general, the models trained for the MT-based recommender are much more portable across domains than those trained for the alignment-based approach. These approaches were compared to a training-free approach (see Section 5), which also uses MT as a source of evidence for word-keeping recommendation, but which does not need any training. The experiments confirm that the results obtained with the training-free MT-based recommender are worse than those obtained with the trained recommender when it is trained on the same TM to be used for translating new texts. However, it is advisable to use the training-free MT-based approach when no recommendation models for the same TM are available for the trained MT-based recommender.

In summary, the MT-based approaches perform better than the alignment-based approaches when accuracy is more important than coverage, or when they are trained on out-of-domain TMs. With regard to the MT-based approaches, it is better to use a trained MT-based recommender when a model is available for the pair of languages and MT system(s) to be used for translating, and to use the training-free MT-based recommender otherwise.

The principal conclusion of this work is that the three approaches are comparable and useful depending on the needs of the translator and the resources available for translation. It might even be possible to combine the three approaches (trained MT-based, training-free MT-based, and statistical alignment-based) in order to prove, for example, recommendations for those words not covered by some of the approaches and not by the others.

The results obtained in this study have also opened up other new horizons for future work, such as: extending the method so as to be able to not only provide the user with recommendations as to which words to keep unedited, but also actively suggest a translation for the words to change; trying alternative parametric classifiers; and using other sources of bilingual knowledge, such as glossaries, dictionaries, bilingual concordancers, etc. to improve the results of the MT-based approaches for word-keeping recommendation.

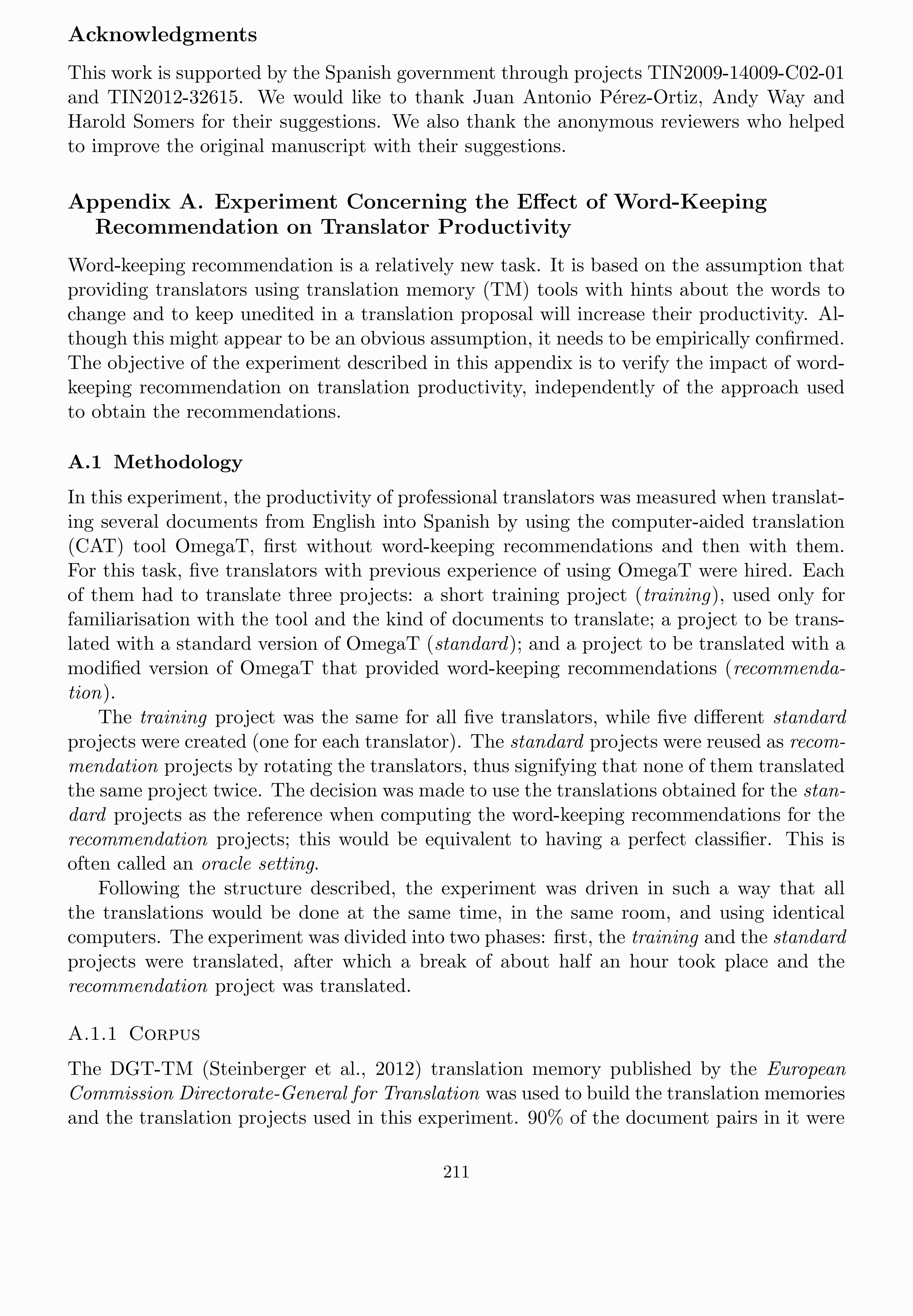
Appendix A presents a study that confirms the usefulness of word-keeping recommendation for translators, showing an improvement in productivity of up to 14% in a translation task from Spanish into English. We plan to extend these experiments to explore new ways of performing word-keeping recommendation. For instance, it would be interesting to compare the productivity of translators when receiving recommendations only for content words, optionally with partial recommendations (on stems), or receiving only change recommendations. We also believe that it would be interesting to evaluate the amount of minimum recommendations needed in a segment to make this tool useful for the translators, by computing the productivity of translators as regards proposals with low coverage. One of our main interests is to be able to model the cost of errors in recommendations, i.e. to confirm whether a wrong “keeping” recommendation is more expensive for a translator than a wrong “changing” recommendation. All these ideas require a new set of experiments with professional translators in order to obtain the optimal method with which to present recommendations, in order to maximise the improvement in productivity already shown in Appendix A.

A study on the impact of noise in the data set used for evaluation in this paper is included in Appendix B. This study uses an heuristic[31] to filter out free or wrong translations in the data sets. The translated materials obtained from the experiments described in Appendix A are additionally used as a clean data set produced directly from professional translators. The results in this appendix show that the accuracy in classification can be significantly improved when using clean data sets.

Finally, a prototype of a plug-in for the free/open-source CAT system OmegaT[32] which implements the training-free approach described in Section 5 as a proof of concept, is available and can be downloaded from `http://www.dlsi.ua.es/~mespla/edithints.html`. This prototype uses free on-line MT systems to perform word-keeping recommendation, thus confirming the technical feasibility of this approach as regards making on-the-fly recommendations in real-world settings.

[31] This heuristic is based on the distance between the segment to translate S' and the source side of the translation proposal S, and the distance between the reference translation T' used as a gold standard for evaluation and the target side of the translation proposal T.

[32] `http://www.omegat.org` [last visit: 15th May 2015]



Acknowledgments

This work is supported by the Spanish government through projects TIN2009-14009-C02-01 and TIN2012-32615. We would like to thank Juan Antonio Pérez-Ortiz, Andy Way and Harold Somers for their suggestions. We also thank the anonymous reviewers who helped to improve the original manuscript with their suggestions.

Appendix A. Experiment Concerning the Effect of Word-Keeping Recommendation on Translator Productivity

Word-keeping recommendation is a relatively new task. It is based on the assumption that providing translators using translation memory (TM) tools with hints about the words to change and to keep unedited in a translation proposal will increase their productivity. Although this might appear to be an obvious assumption, it needs to be empirically confirmed. The objective of the experiment described in this appendix is to verify the impact of word-keeping recommendation on translation productivity, independently of the approach used to obtain the recommendations.

A.1 Methodology

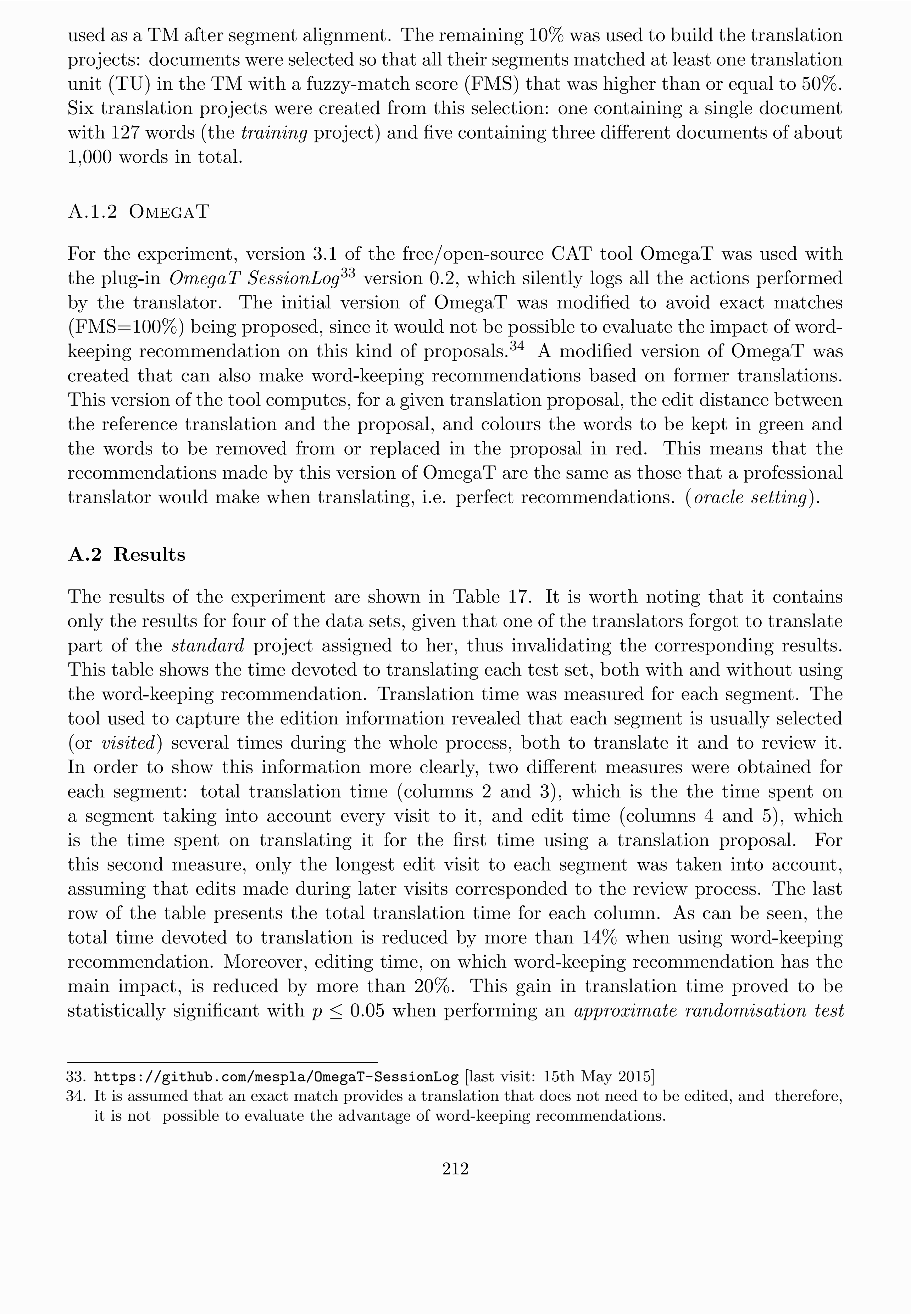
In this experiment, the productivity of professional translators was measured when translating several documents from English into Spanish by using the computer-aided translation (CAT) tool OmegaT, first without word-keeping recommendations and then with them. For this task, five translators with previous experience of using OmegaT were hired. Each of them had to translate three projects: a short training project (*training*), used only for familiarisation with the tool and the kind of documents to translate; a project to be translated with a standard version of OmegaT (*standard*); and a project to be translated with a modified version of OmegaT that provided word-keeping recommendations (*recommendation*).

The *training* project was the same for all five translators, while five different *standard* projects were created (one for each translator). The *standard* projects were reused as *recommendation* projects by rotating the translators, thus signifying that none of them translated the same project twice. The decision was made to use the translations obtained for the *standard* projects as the reference when computing the word-keeping recommendations for the *recommendation* projects; this would be equivalent to having a perfect classifier. This is often called an *oracle setting.*

Following the structure described, the experiment was driven in such a way that all the translations would be done at the same time, in the same room, and using identical computers. The experiment was divided into two phases: first, the *training* and the *standard* projects were translated, after which a break of about half an hour took place and the *recommendation* project was translated.

A.1.1 Corpus

The DGT-TM (Steinberger et al., 2012) translation memory published by the *European Commission Directorate-General for Translation* was used to build the translation memories and the translation projects used in this experiment. 90% of the document pairs in it were



used as a TM after segment alignment. The remaining 10% was used to build the translation projects: documents were selected so that all their segments matched at least one translation unit (TU) in the TM with a fuzzy-match score (FMS) that was higher than or equal to 50%. Six translation projects were created from this selection: one containing a single document with 127 words (the *training* project) and five containing three different documents of about 1,000 words in total.

A.1.2 OmegaT

For the experiment, version 3.1 of the free/open-source CAT tool OmegaT was used with the plug-in *OmegaT SessionLog*[33] version 0.2, which silently logs all the actions performed by the translator. The initial version of OmegaT was modified to avoid exact matches (FMS=100%) being proposed, since it would not be possible to evaluate the impact of word-keeping recommendation on this kind of proposals.[34] A modified version of OmegaT was created that can also make word-keeping recommendations based on former translations. This version of the tool computes, for a given translation proposal, the edit distance between the reference translation and the proposal, and colours the words to be kept in green and the words to be removed from or replaced in the proposal in red. This means that the recommendations made by this version of OmegaT are the same as those that a professional translator would make when translating, i.e. perfect recommendations. (*oracle setting*).

A.2 Results

The results of the experiment are shown in Table 17. It is worth noting that it contains only the results for four of the data sets, given that one of the translators forgot to translate part of the *standard* project assigned to her, thus invalidating the corresponding results. This table shows the time devoted to translating each test set, both with and without using the word-keeping recommendation. Translation time was measured for each segment. The tool used to capture the edition information revealed that each segment is usually selected (or *visited*) several times during the whole process, both to translate it and to review it. In order to show this information more clearly, two different measures were obtained for each segment: total translation time (columns 2 and 3), which is the the time spent on a segment taking into account every visit to it, and edit time (columns 4 and 5), which is the time spent on translating it for the first time using a translation proposal. For this second measure, only the longest edit visit to each segment was taken into account, assuming that edits made during later visits corresponded to the review process. The last row of the table presents the total translation time for each column. As can be seen, the total time devoted to translation is reduced by more than 14% when using word-keeping recommendation. Moreover, editing time, on which word-keeping recommendation has the main impact, is reduced by more than 20%. This gain in translation time proved to be statistically significant with $p \leq 0.05$ when performing an *approximate randomisation test*

[33] `https://github.com/mespla/OmegaT-SessionLog` [last visit: 15th May 2015]

[34] It is assumed that an exact match provides a translation that does not need to be edited, and therefore, it is not possible to evaluate the advantage of word-keeping recommendations.

test set	**total time**		**edition time**	
	without WKR	**with WKR**	**without WKR**	**with WKR**
1	3,664s	2,611s	2,441s	1,917s
2	3,613s	3,467s	3,080s	2,293s
3	4,251s	3,709s	2,674s	2,310s
4	3,787s	3,315s	2,432s	1,937s
all test sets	15,315s	**13,102s**	10,627s	**8,457**s

Table 17: Time spent on translation. Columns 2 and 3 compare the total time spent translating each test set, respectively, using the version of OmegaT without and with word-keeping recommendation. Columns 4 and 5 present the same comparison, but only taking into account the time actually spent reviewing the test set.

with 1,000 iterations.[35] The free/open-source tool SIGF V.2 by Padó (2006) was used for these experiments.

The results obtained in this experiment confirm the assumption that word-keeping recommendation can significantly improve the productivity of translators who use translation memory tools. Although a more extensive experiment, including more translators and documents from other domains, would be needed to confirm this, current results are encouraging. In addition, all the translators participating in the experiment agreed that word-keeping recommendation is useful for translators when working with TM-based CAT tools.

It is worth noting that the experimental framework presented in this appendix has been specifically designed to measure word-keeping recommendation and the results obtained here cannot therefore be straightforwardly assumed for every translation project. For example, the projects translated in this experiment used a TM, thus ensuring that at least one translation proposal would be provided with an FMS that was higher than 50%, but a TM with this type of coverage may not be available for a given project. In addition, translations performed by humans were used in this experiment to compute the word-keeping recommendations, in what would usually be called a *gold standard*. These translations would obviously not be available in a real scenario and recommendations would be approximate. The use of gold-standard-based recommendations may also boost the confidence of the translators when using the tool, since in this experiment it was correct most of the times. We can therefore consider that these results correspond to an upper bound in productivity gain. Nevertheless, the results obtained in this experiment have allowed us to obtain a clearer idea of the usefulness of word-keeping recommendation and confirm the relevance of the problem of obtaining fast and accurate word-keeping recommendations.

[35] Here, approximate randomisation is applied to the time devoted to translating each segment with and without word-keeping recommendation in the concatenated data sets. This method first computes the difference in time needed to translate the entire data sets. It then randomly interchanges the time spent translating some of these segments between both sets of results and recomputes this total time. If an equal or higher time gain can be obtained with these randomised lists of times, this means that the result is not significant.

Appendix B. Experiments with High Quality *Gold Standards*

In this appendix we tackle the problem associated with the use of free translations as references for evaluation. As already mentioned in Section 7, for a pair of segments (S'_l, T'_l) in our test set, we obtain all the matching TUs (S_i, T_i), and a set of word-keeping recommendations that are provided for every segment T_i. T'_l is then used as a gold standard for these recommendations for the purpose of evaluation. This method assumes that the way in which S'_l is translated into T'_l is similar to the way in which S_i is translated into T_i, thus enabling the use of T'_l as a reference.[36] However, this may not be the case for several reasons, such as wrong segment alignments, errors in translations, or, in our case, free (but still adequate) translations. Following the example shown in Section 4, we illustrate the impact of a free translation on our evaluation method. Let us assume that the segment S' to be translated is *"la situación política parece ser difícil"*, that a matching TU (S, T) retrieved by the CAT tool is *("la situación humanitaria parece ser difícil", "the humanitarian situation appears to be difficult"),* and that the gold standard T' in the test set is *"the situation, from a political point of view, appears tortuous"*, which is a semantically valid translation of S', but very different to T. When checking the validity of the translation as occurred in Section 4, the only words common to T' and T are *the*, *appears*, and *situation*, and the remaining words would be considered as words to change in order to produce a correct translation. However, it is sufficient to replace the word *humanitarian* with *political* in T to produce a valid translation of S'. It is therefore obvious that T' is not a good reference with which to evaluate the recommendations performed on T.

As a consequence of the free translations in our test set, a fraction of the recommendations which are actually correct are considered inadequate during the evaluation since they do not match the reference in the test set. The accuracy obtained for all the approaches presented in this work is therefore lower than expected.

Although the loss of accuracy affects all the methods in this work in the same way and the conclusions that are obtained are therefore valid, we wished to see if more reliable results as regards the performance of these approaches could be attained. We therefore performed a set of additional experiments in order to bypass the problem of the free translations. On the one hand, we repeated some of the experiments shown in Section 7 but by using a constrained test set in which all those pairs of segments which were likely to be wrong (or free) translations were discarded. On the other hand, we performed an experiment by re-using the test set and the TMs described in Appendix A.

As mentioned previously, in the first group of experiments we defined a constraint in order to attempt to evaluate only those pairs of segments from the test set in Section 6.2 which are more reliable. This was done by employing a filtering based on the fuzzy-match score (FMS) used to choose the candidate TUs for a given segment to be translated. This condition relies on the assumption that the FMS between S and S' (FMS$_S$) should be similar to the FMS between T and T' (FMS$_T$), since the number of words that differed in both pairs of segments should be proportional for both languages. Based on this idea, we set a threshold ϕ so that only pairs of TUs fulfilling the condition $|\text{FMS}_S - \text{FMS}_T| \leq \phi$ were

[36] By *similar* we mean that the matching parts between S'_l and S_i are translated in the same way, thus producing differences between T'_l and T_i only in those parts corresponding to the differences between S'_l and S_i.

$\Theta(\%)$	**domain**	**no filtering**		$\phi \leq 0.05$	
		TU_{avg}	N_{words}	TU_{avg}	N_{words}
≥ 60	02.40.10.40	3.71	95,881	1.59	44,240
	05.20.20.10	0.62	9,718	0.39	6,198
	06.30.10.00	5.68	34,339	0.52	6,956
≥ 70	02.40.10.40	2.36	65,865	1.16	31,022
	05.20.20.10	0.37	6,883	0.26	4,862
	06.30.10.00	0.99	10,327	0.26	4,194
≥ 80	02.40.10.40	1.58	46,519	0.97	24,928
	05.20.20.10	0.14	3,015	0.09	1,889
	06.30.10.00	0.45	4,726	0.09	2,143
≥ 90	02.40.10.40	0.70	26,625	0.50	15,154
	05.20.20.10	0.05	1,599	0.03	975
	06.30.10.00	0.03	1,268	0.03	943

Table 18: Average number of matching TUs (TU_{avg}) per segment and total number of words (N_{words}) for which to provide a recommendation when translating `es`→`en` for the three different domains. The results were obtained for different ranges of the FMS threshold (Θ), both when filtering with $\phi = 0.05$ and when no filtering was applied.

used for both training and testing. It is worth mentioning that some experiments were also performed by applying this filtering only to the test set, but the difference in the results was not significant. For our experiments, we arbitrarily set the value of ϕ to 0.05, i.e. a divergence of 5% was permitted between the FMS of the source language segments and that of the target language segments, since it is a threshold that constrains the examples used in a highly controlled scenario, but a reasonable number of samples is maintained for our experiments, as shown in Table 18.[37]

B.1 Experiments with Constrained Test Sets

This table shows, for the `es`→`en` language pair and for fuzzy-match scores in four different ranges, the average number of TUs matched per segment to be translated and the total number of words for which a recommendation should be provided. The results were obtained both when filtering with threshold ϕ and when no filtering was applied. It is worth noting that in the case of domains 05.20.20.10 and 06.30.10.00 there were noticeable differences in matching when no restriction was applied, while more similar data was obtained when filtering with $\phi = 0.05$. This has led us to believe that the TUs belonging to domain 05.20.20.10 are more regular in translation than those in domain 06.30.10.00. As will be

[37] Note that the objective of these experiments is not to compare the different approaches (this has been already done), but rather to confirm whether an improvement in accuracy exists when using less noisy data sets. The statistical significance between the different approaches has not therefore been re-computed.

Θ(%)	Method	A (%)	NC (%)	P_{k} (%)	R_{k} (%)	P_{c} (%)	R_{c} (%)
≥ 60	**FMS-only**	84.7±.3	100%±0	84.7±.3	100%±0	—	0%±0
	alignment	96.4±.2	4.8±.2	98.2±.1	97.5±.2	85.8±.3	89.4±.3
	trained	96.9±.2	4.7±.2	97.9±.1	98.4±.1	90.8±.3	88.2±.3
	training-free	95.2±.2	4.7±.2	96.5±.2	97.9±.1	87.3±.3	79.9±.4
≥ 70	**FMS-only**	90.5±.3	100%±0	90.5±.3	100%±0	—	0%±0
	alignment	97.0±.2	4.7±.2	98.5±.1	98.2±.2	81.7±.4	83.9±.4
	trained	97.4±.2	4.4±.2	98.3±.2	98.8±.1	87.3±.4	82.7±.4
	training-free	96.1±.2	4.4±.2	97.1±.2	98.6±.1	83.4±.4	69.8±.5
≥ 80	**FMS-only**	93.2±.3	100%±0	93.2±.3	100%±0	—	0%±0
	alignment	97.4±.2	4.6±.3	98.7±.1	98.5±.2	77.1±.5	79.4±.5
	trained	98.0±.2	4.6±.3	98.7±.2	99.2±.1	86.5±.4	79.5±.5
	training-free	96.7±.2	4.6±.3	97.5±.2	98.0±.2	79.8±.5	61.7±.6
≥ 90	**FMS-only**	96.5±.3	100%±0	96.5±.3	100%±0	—	0%±0
	alignment	97.9±.2	4.1±.3	98.9±.2	98.9±.2	68.0±.8	68.0±.8
	trained	98.6±.2	4.2±.3	99.0±.2	99.6±.1	81.7±.6	62.8±.8
	training-free	98.3±.2	4.2±.3	98.9±.2	99.4±.1	74.0±.7	60.8±.8

Table 19: Comparison of the results obtained using the trained MT-based approach, the training-free MT-based approach, the alignment-based approach and the naïve FMS-only baseline. The accuracy (A) and fraction of words not covered (NC) are reported, together with the precision (P) and recall (R) for both keeping recommendations and changing recommendations. The results were obtained for several values of the FMS threshold Θ when translating es→en in domain 02.40.10.40, using all the MT systems available and filtering with $\phi = 0.05$ (see text).

observed, with this threshold, approximately half of the training samples are kept for domain 02.40.10.40 and about two thirds for domain 05.20.20.10. The case of domain 06.30.10.00 is different; the filtering removes far more training samples for low values of the FMS threshold Θ, while for higher values the loss is not so high, and similar to that of domain 05.20.20.10.

Table 19 is the equivalent of Table 12, which contains a detailed comparison of all the approaches, but using the filtering described above on the data set. As will be noted, the results obtained in this case are clearly better for all the approaches than those obtained in the experiments with no filtering.

Finally, Table 20 shows the accuracy obtained by both the trained MT-based approach and the alignment-based approach for all language pairs, as occurs in Table 10. It is worth noting that, although the differences between the results obtained with both approaches are similar, all of them are noticeably better.

B.2 Experiment With Human-Produced Test Sets

In this second group of experiments we used the documents described in Appendix A as a test set to evaluate word-keeping recommendation. In this case, the original documents in

	Θ ≥ 60(%)		Θ ≥ 70(%)		Θ ≥ 80(%)		Θ ≥ 90(%)	
lang. pair	trained	alignment	trained	alignment	trained	alignment	trained	alignment
es→en	96.9±.2	96.4±.2	97.5±.2	97.0±.2	98.0±.2	97.4±.2	98.5±.2	97.9±.2
en→es	95.1±.2	93.6±.2	96.4±.2	94.8±.2	97.1±.2	95.5±.2	97.8±.2	96.9±.3
de→en	96.9±.2	96.3±.2	97.7±.2	96.8±.2	98.3±.2	97.4±.2	98.3±.2	97.5±.3
en→de	96.3±.2	94.8±.2	97.2±.2	95.7±.2	97.6±.2	96.2±.2	97.9±.2	97.0±.3
fr→en	96.9±.2	96.7±.2	97.9±.2	97.7±.2	98.4±.2	98.0±.2	98.5±.2	98.1±.2
en→fr	95.9±.2	95.5±.2	97.4±.2	96.8±.2	98.1±.1	97.3±.2	98.3±.2	97.5±.2
fi→en	96.0±.2	96.0±.2	97.3±.2	97.1±.2	97.7±.2	97.5±.2	98.0±.3	97.5±.3
en→fi	96.3±.2	94.8±.3	97.2±.2	95.5±.3	97.7±.2	96.0±.3	97.7±.3	97.5±.3
es→fr	95.6±.2	95.3±.2	96.8±.2	96.5±.2	97.7±.2	97.2±.2	98.1±.2	97.4±.2
fr→es	95.2±.2	94.8±.2	96.7±.2	96.3±.2	97.3±.2	97.0±.2	97.4±.2	97.0±.3

Table 20: Accuracy (A) obtained with both the trained MT-based approach and the alignment-based approach when translating between all the language pairs in domain 02.40.10.40. The results were obtained for several values of the FMS threshold Θ and using all available MT systems for each language pair.

Θ(%)	A (%)	NC (%)
≥ 60	97.8±.1	10.0±.2
≥ 70	98.6±.1	8.5±.2
≥ 80	99.0±.1	8.1±.3
≥ 90	98.7±.2	7.3±.4

Θ(%)	A (%)	NC (%)
≥ 60	95.6±.1	9.8±.2
≥ 70	96.2±.1	8.5±.2
≥ 80	96.7±.2	8.1±.2
≥ 90	96.4±.2	8.1±.3

Table 21: Accuracy (A) and fraction of words not covered (NC) obtained when translating with the trained MT-based approach by reusing the data set described in Appendix A. The left-hand table contains the results when translating Spanish into English, while the right-hand table contains the results when translating English into Spanish.

Spanish were translated into English by professional translators, who were told to translate them as faithfully as possible. These parallel documents were therefore expected to totally fit the requirements of the evaluation.

In this experiment, the TM used by the professional translators in Appendix A was used to evaluate the translation of the texts from English into Spanish and vice versa. In-domain models were also trained on this TM which, as already mentioned above, consists of only 629 TUs. Table 21 presents the accuracy and the fraction of words not covered that were obtained for this data set, for en→es and for es→en. Although the coverage is slightly lower than that obtained by the system with other data sets, the accuracy is much better, and is even better than that obtained with the constrained test sets.

The results presented in this appendix allow us to confirm that the accuracy of the approaches presented in this work may be noticeably higher than those presented in Section 7, but the lack of a valid gold standard for our experiments only allows us to approximate these results.

References

Ahrenberg, L., Andersson, M., & Merkel, M. (2000). *Parallel text processing: alignment and use of translation corpora*, chap. A knowledge-lite approach to word alignment. Kluwer Academic Publishers. Edited by J. Véronis.

Bergstra, J. S., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems 24*, pp. 2546–2554. Curran Associates, Inc.

Bertoldi, N., Farajian, A., & Federico, M. (2009). Online word alignment for online adaptive machine translation. In *Proceedings of the Workshop on Humans and Computer-assisted Translation*, pp. 120–127, Gothenburg, Sweden.

Biçici, E., & Dymetman, M. (2008). Dynamic translation memory: Using statistical machine translation to improve translation memory fuzzy matches. In *Proceedings of the 9th International Conference on Intelligent Text Processing and Computational Linguistics*, Vol. 4919 of *LNCS*, pp. 454–465, Haifa, Israel.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., & Tamchyna, A. (2014). Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 12–58, Baltimore, USA.

Bourdaillet, J., Huet, S., Langlais, P., & Lapalme, G. (2010). TransSearch: from a bilingual concordancer to a translation finder. *Machine Translation*, *24*(3–4), 241–271.

Bowker, L. (2002). *Computer-aided translation technology: a practical introduction*, chap. Translation-memory systems, pp. 92–127. University of Ottawa Press.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, *19*(2), 263–311.

de Gispert, A., Blackwood, G., Iglesias, G., & Byrne, W. (2013). N-gram posterior probability confidence measures for statistical machine translation: an empirical study. *Machine Translation*, *27*(2), 85–114.

Depraetere, I. (2008). LEC Power Translator 12. *MultiLingual*, *September 2008*, 18–22.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification* (second edition). John Wiley and Sons Inc.

Esplà, M., Sánchez-Martínez, F., & Forcada, M. L. (2011). Using word alignments to assist computer-aided translation users by marking which target-side words to change or keep unedited. In *Proceedings of the 15th Annual Conference of the European Associtation for Machine Translation*, pp. 81–89, Leuven, Belgium.

Esplà-Gomis, M., Sánchez-Martínez, F., & Forcada, M. L. (2011). Using machine translation in computer-aided translation to suggest the target-side words to change. In *Proceedings of the 13th Machine Translation Summit*, pp. 172–179, Xiamen, China.

Esplà-Gomis, M., Sánchez-Martínez, F., & Forcada, M. L. (2015). Using on-line available sources of bilingual information for word-level machine translation quality estimation. In *Proceedings of the 18th Annual Conference of the European Associtation for Machine Translation*, pp. 19–26, Antalya, Turkey.

Esplà-Gomis, M., Sánchez-Martínez, F., & Forcada, M. L. (2012a). Using external sources of bilingual information for on-the-fly word alignment. Tech. rep., Universitat d'Alacant.

Esplà-Gomis, M., Sánchez-Martínez, F., & Forcada, M. L. (2012b). A simple approach to use bilingual information sources for word alignment. *Procesamiento de Lenguaje Natural*, *49*.

European Commission Directorate-General for Translation (2009). *Translation Tools and Workflow*. Directorate-General for Translation of the European Commission.

European Commission Joint Research Center (2007). EUR-Lex pre-processing. `http://optima.jrc.it/Resources/Documents/DGT-TM_EUR-LEX-preprocessing.pdf`. Last retrieved: 15th May 2015.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., & Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, *25*(2), 127–144. Special Issue on Free/Open-Source Machine Translation.

Foster, G., & Kuhn, R. (2007). Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pp. 128–135, Prague, Czech Republic.

Free Software Fundation (2007). GNU general public license, version 3. `http://www.gnu.org/licenses/gpl.html`. Last retrieved: 15th May 2015.

Gao, Q., Lewis, W., Quirk, C., & Hwang, M. (2011). Incremental training and intentional over-fitting of word alignment. In *Proceedings of the 13th Machine Translation Summit*, pp. 106–113, Xiamen, China.

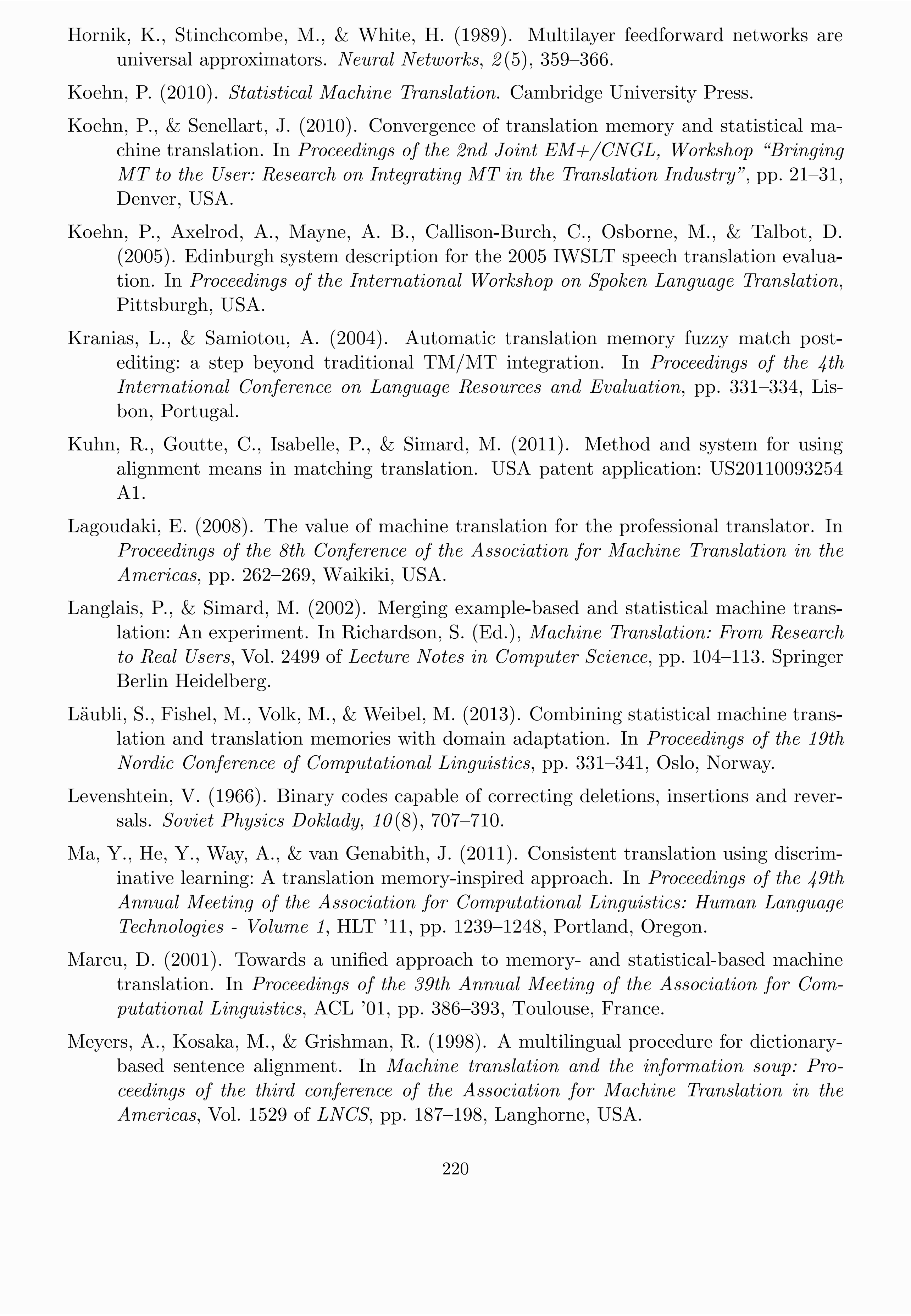
Gao, Q., & Vogel, S. (2008). Parallel implementations of word alignment tool. In *Proceedings of the Software Engineering, Testing, and Quality Assurance for Natural Language Processing Workshop*, pp. 49–57, Columbus, USA.

Garcia, I. (2005). Long term memories: Trados and TM turn 20. *Journal of Specialised Translation*, *4*, 18–31.

Garcia, I. (2012). Machines, translations and memories: language transfer in the web browser. *Perspectives*, *20*(4), 451–461.

Gough, N., Way, A., & Hearne, M. (2002). Example-based machine translation via the web. In Richardson, S. D. (Ed.), *Machine Translation: From Research to Real Users*, Vol. 2499 of *Lecture Notes in Computer Science*, pp. 74–83. Springer Berlin Heidelberg.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: an Update. *SIGKDD Explorations*, *11*(1), 10–18.



Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366.

Koehn, P. (2010). *Statistical Machine Translation.* Cambridge University Press.

Koehn, P., & Senellart, J. (2010). Convergence of translation memory and statistical machine translation. In *Proceedings of the 2nd Joint EM+/CNGL, Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry"*, pp. 21–31, Denver, USA.

Koehn, P., Axelrod, A., Mayne, A. B., Callison-Burch, C., Osborne, M., & Talbot, D. (2005). Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation*, Pittsburgh, USA.

Kranias, L., & Samiotou, A. (2004). Automatic translation memory fuzzy match post-editing: a step beyond traditional TM/MT integration. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pp. 331–334, Lisbon, Portugal.

Kuhn, R., Goutte, C., Isabelle, P., & Simard, M. (2011). Method and system for using alignment means in matching translation. USA patent application: US20110093254 A1.

Lagoudaki, E. (2008). The value of machine translation for the professional translator. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas*, pp. 262–269, Waikiki, USA.

Langlais, P., & Simard, M. (2002). Merging example-based and statistical machine translation: An experiment. In Richardson, S. (Ed.), *Machine Translation: From Research to Real Users*, Vol. 2499 of *Lecture Notes in Computer Science*, pp. 104–113. Springer Berlin Heidelberg.

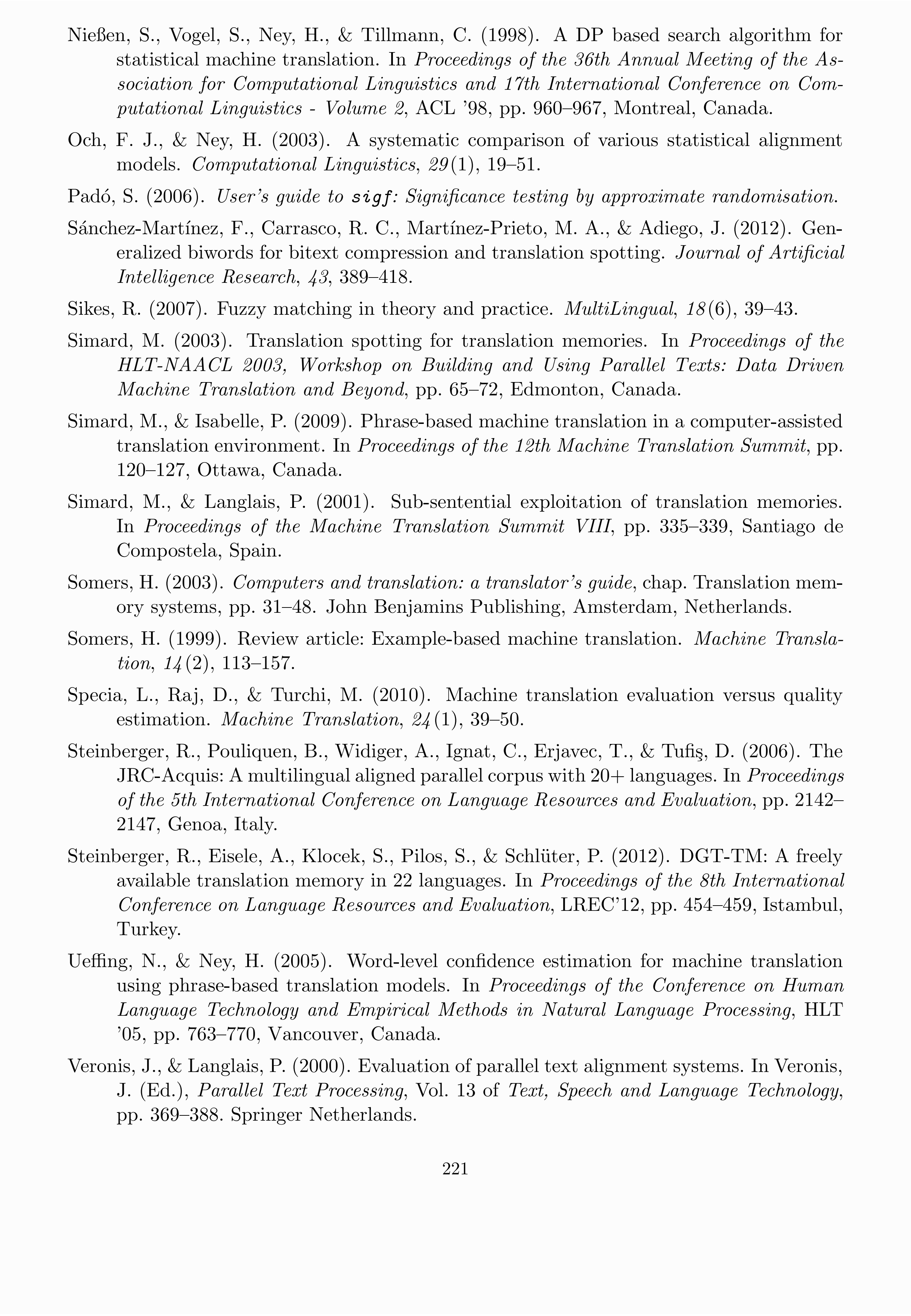
Läubli, S., Fishel, M., Volk, M., & Weibel, M. (2013). Combining statistical machine translation and translation memories with domain adaptation. In *Proceedings of the 19th Nordic Conference of Computational Linguistics*, pp. 331–341, Oslo, Norway.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Ma, Y., He, Y., Way, A., & van Genabith, J. (2011). Consistent translation using discriminative learning: A translation memory-inspired approach. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pp. 1239–1248, Portland, Oregon.

Marcu, D. (2001). Towards a unified approach to memory- and statistical-based machine translation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, ACL '01, pp. 386–393, Toulouse, France.

Meyers, A., Kosaka, M., & Grishman, R. (1998). A multilingual procedure for dictionary-based sentence alignment. In *Machine translation and the information soup: Proceedings of the third conference of the Association for Machine Translation in the Americas*, Vol. 1529 of *LNCS*, pp. 187–198, Langhorne, USA.



Nießen, S., Vogel, S., Ney, H., & Tillmann, C. (1998). A DP based search algorithm for statistical machine translation. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, pp. 960–967, Montreal, Canada.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19–51.

Padó, S. (2006). *User's guide to `sigf`: Significance testing by approximate randomisation.*

Sánchez-Martínez, F., Carrasco, R. C., Martínez-Prieto, M. A., & Adiego, J. (2012). Generalized biwords for bitext compression and translation spotting. *Journal of Artificial Intelligence Research*, *43*, 389–418.

Sikes, R. (2007). Fuzzy matching in theory and practice. *MultiLingual*, *18*(6), 39–43.

Simard, M. (2003). Translation spotting for translation memories. In *Proceedings of the HLT-NAACL 2003, Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pp. 65–72, Edmonton, Canada.

Simard, M., & Isabelle, P. (2009). Phrase-based machine translation in a computer-assisted translation environment. In *Proceedings of the 12th Machine Translation Summit*, pp. 120–127, Ottawa, Canada.

Simard, M., & Langlais, P. (2001). Sub-sentential exploitation of translation memories. In *Proceedings of the Machine Translation Summit VIII*, pp. 335–339, Santiago de Compostela, Spain.

Somers, H. (2003). *Computers and translation: a translator's guide*, chap. Translation memory systems, pp. 31–48. John Benjamins Publishing, Amsterdam, Netherlands.

Somers, H. (1999). Review article: Example-based machine translation. *Machine Translation*, *14*(2), 113–157.

Specia, L., Raj, D., & Turchi, M. (2010). Machine translation evaluation versus quality estimation. *Machine Translation*, *24*(1), 39–50.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., & Tufiş, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pp. 2142–2147, Genoa, Italy.

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., & Schlüter, P. (2012). DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, LREC'12, pp. 454–459, Istambul, Turkey.

Ueffing, N., & Ney, H. (2005). Word-level confidence estimation for machine translation using phrase-based translation models. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pp. 763–770, Vancouver, Canada.

Veronis, J., & Langlais, P. (2000). Evaluation of parallel text alignment systems. In Veronis, J. (Ed.), *Parallel Text Processing*, Vol. 13 of *Text, Speech and Language Technology*, pp. 369–388. Springer Netherlands.

Vogel, S., Ney, H., & Tillmann, C. (1996). HMM-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics*, pp. 836–841, Copenhagen, Denmark.

Zhechev, V., & van Genabith, J. (2010). Seeding statistical machine translation with translation memory output through tree-based structural alignment. In *Proceedings of the COLING'10, Workshop on Syntax and Structure in Statistical Translation*, pp. 43–51, Beijing, China.

Chapter 3

Methods for word-level quality estimation in machine translation using external sources of bilingual information

This chapter describes the research conducted on the use of external SBI to estimate the quality at the word level in MT. This research is based on the ideas in the previous chapter, in which these techniques were applied to CAT. The underlying working hypothesis in this chapter is:

Hypothesis #4: it is possible to take the SBI-based methods for word-level QE in TM-based CAT and adapt them for their use in MT.

The problem of MT is conceptually similar to that of TM-based CAT when seen from a word-level QE perspective: given a SL segment to be translated, a translation hypothesis is proposed which is likely to need post-editing to obtain an adequate translation. However, there is an important difference: in CAT, the translation suggestion T for a given SL segment S' is usually well formed, since it is an adequate translation of another segment S presumably performed by a professional translator. Therefore, the problem of word-level QE in CAT is to detect which parts of T are adequate to produce a translation of S'. However, in the case of MT, the translation hypothesis is specifically produced for S', although it is likely to be partially inadequate. In this case, the task of word-level QE in MT consists in detecting which words of the translation hypothesis are adequate and which are inadequate.

When trying to adapt the method defined for TM-based CAT to MT, one encounters another important difference: in CAT, the objective was to project information obtained by a monolingual comparison between the SL segments S and S' onto the

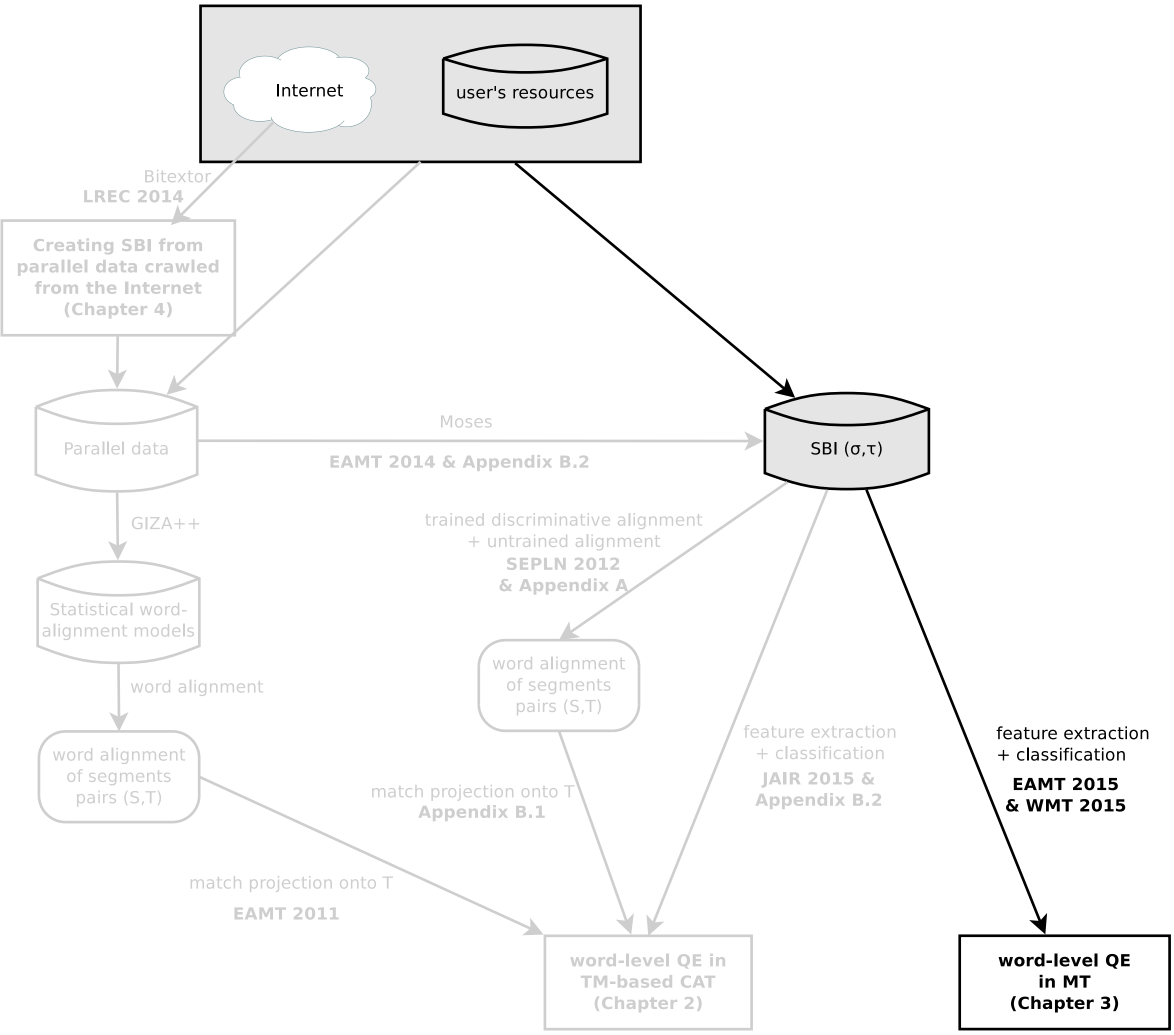

Internet
user's resources
Bitextor
LREC 2014
Creating SBI from parallel data crawled from the Internet (Chapter 4)
Parallel data
Moses
EAMT 2014 & Appendix B.2
SBI (σ,τ)
GIZA++
Statistical word-alignment models
word alignment
word alignment of segments pairs (S,T)
trained discriminative alignment + untrained alignment
SEPLN 2012 & Appendix A
word alignment of segments pairs (S,T)
match projection onto T
Appendix B.1
feature extraction + classification
JAIR 2015 & Appendix B.2
feature extraction + classification
EAMT 2015 & WMT 2015
match projection onto T
EAMT 2011
word-level QE in TM-based CAT (Chapter 2)
word-level QE in MT (Chapter 3)

Figure 3.1: This diagram highlights the research being reported in Chapter 3 (and the publications concerned) on the development of word-level QE methods in MT that are able to use SBI, and relates it to the rest of the work reported in this dissertation.

TL segment T, while in MT, this kind of information does not exist. Therefore, a fully cross-lingual strategy is required in this case, in which SBI are used as partial pseudo-references. The research conducted in the creation of this new strategy is highlighted and put in context in Figure 3.1.

A detailed description of the new approaches developed is available in the following two papers included in this chapter:

- Esplà-Gomis, M., Sánchez-Martínez, F., and Forcada, M.L. 2015. Using on-line available sources of bilingual information for word-level machine translation quality estimation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, p. 19–26, Antalya, Turkey, May 11–13, 2015. **[Reprinted publication 3.1]**
- Esplà-Gomis, M., Sánchez-Martínez, F., and Forcada, M.L. 2015. UAlacant word-level machine translation quality estimation system at WMT 2015. In *Proceedings of the 10th Workshop on Statistical Machine Translation*, p. 309–315, Lisbon, Portugal, September 17–18, 2015. **[Reprinted publication 3.2]**

Reprinted publication 3.1 describes the method proposed and the collection of features used by a binary classifier, which are divided into positive features, that provide evidence that a word is adequate, and negative features, that provide evidence that a word is inadequate. In total, the proposed approach uses about 70 features, which is reasonably small when compared to some other approaches (Camargo de Souza et al., 2014; Biçici and Way, 2014).

To evaluate the approach in reprinted publication 3.1 we used one of the most recent datasets available at the time of publishing it: the one provided for the shared task on word-level QE in MT in the 2014 edition of the Workshop on Statistical Machine Translation (Bojar et al., 2014). This data set provides an evaluation framework for two language pairs: English–German and Spanish–English, in both translation directions. The results confirm the feasibility of the approach and, for some translation directions, show that its performance is comparable to that of the best systems competing in the task.

Reprinted publication 3.2 describes the setting of our approach for the shared task on word-level QE in MT in the 2015 edition of the Workshop on Statistical Machine Translation (Bojar et al., 2015); in this case, evaluation data was only provided for Spanish–English. Two systems were submitted to the task: one using only the features described in reprinted publication 3.1 and another combining them with the collection of baseline features provided by the organisers of the task. The approach using only the features described in reprinted publication 3.1 ranked third, while that combining them with the baseline features ranked first (Bojar et al., 2015).

Using on-line available sources of bilingual information for word-level machine translation quality estimation

Miquel Esplà-Gomis Felipe Sánchez-Martínez Mikel L. Forcada
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain
{mespla,fsanchez,mlf}@dlsi.ua.es

Abstract

This paper explores the use of external sources of bilingual information available on-line for word-level machine translation quality estimation (MTQE). These sources of bilingual information are used as a *black box* to spot sub-segment correspondences between a source-language (SL) sentence S to be translated and a given translation hypothesis T in the target-language (TL). This is done by segmenting both S and T into overlapping sub-segments of variable length and translating them into the TL and the SL, respectively, using the available bilingual sources of information *on the fly*. A collection of features is then obtained from the resulting sub-segment translations, which is used by a binary classifier to determine which target words in T need to be post-edited.

Experiments are conducted based on the data sets published for the word-level MTQE task in the 2014 edition of the Workshop on Statistical Machine Translation (WMT 2014). The sources of bilingual information used are: machine translation (Apertium and Google Translate) and the bilingual concordancer Reverso Context. The results obtained confirm that, using less information and fewer features, our approach obtains results comparable to those of state-of-the-art approaches, and even outperform them in some data sets.

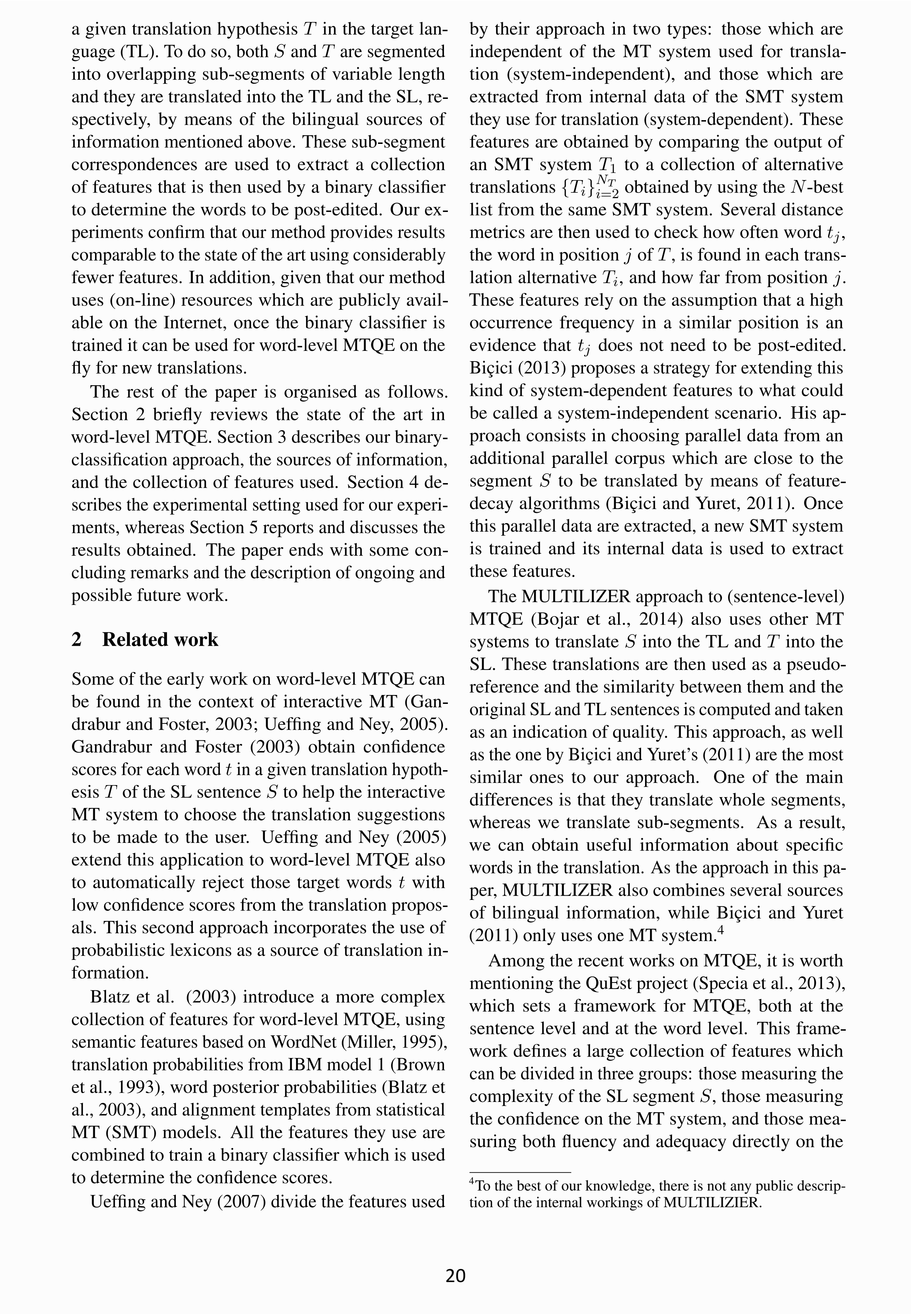
1 Introduction

Recent advances in the field of machine translation (MT) have led to the adoption of this technology by many companies and institutions all around the world in order to bypass the linguistic barriers and reach out to broader audiences. Unfortunately, we are still far from the point of having MT systems able to produce translations with the level of quality required for dissemination in formal scenarios, where human supervision and MT post-editing are unavoidable. It therefore becomes critical to minimise the cost of this human post-editing. This has motivated a growing interest in the field of MT quality estimation (Blatz et al., 2004; Specia et al., 2010; Specia and Soricut, 2013), which is the field that focuses on developing techniques that allow to estimate the quality of the translation hypotheses produced by an MT system.

Most efforts in MT quality estimation (MTQE) are aimed at evaluating the quality of whole translated segments, in terms of post-editing time, number of editions needed, and other related metrics (Blatz et al., 2004). Our work is focused on the sub-field of *word-level MTQE*. The main advantage of word-level MTQE is that it allows not only to estimate the effort needed to post-edit the output of an MT system, but also to guide post-editors on which words need to be post-edited.

In this paper we describe a novel method which uses black-box bilingual resources from the Internet for word-level MTQE. Namely, we combine two on-line MT systems, Apertium[1] and Google Translate,[2] and the bilingual concordancer Reverso Context[3] to spot sub-segment correspondences between a sentence S in the source language (SL) and

[1] http://www.apertium.org
[2] http://translate.google.com
[3] http://context.reverso.net/translation/



a given translation hypothesis T in the target language (TL). To do so, both S and T are segmented into overlapping sub-segments of variable length and they are translated into the TL and the SL, respectively, by means of the bilingual sources of information mentioned above. These sub-segment correspondences are used to extract a collection of features that is then used by a binary classifier to determine the words to be post-edited. Our experiments confirm that our method provides results comparable to the state of the art using considerably fewer features. In addition, given that our method uses (on-line) resources which are publicly available on the Internet, once the binary classifier is trained it can be used for word-level MTQE on the fly for new translations.

The rest of the paper is organised as follows. Section 2 briefly reviews the state of the art in word-level MTQE. Section 3 describes our binary-classification approach, the sources of information, and the collection of features used. Section 4 describes the experimental setting used for our experiments, whereas Section 5 reports and discusses the results obtained. The paper ends with some concluding remarks and the description of ongoing and possible future work.

2 Related work

Some of the early work on word-level MTQE can be found in the context of interactive MT (Gandrabur and Foster, 2003; Ueffing and Ney, 2005). Gandrabur and Foster (2003) obtain confidence scores for each word t in a given translation hypothesis T of the SL sentence S to help the interactive MT system to choose the translation suggestions to be made to the user. Ueffing and Ney (2005) extend this application to word-level MTQE also to automatically reject those target words t with low confidence scores from the translation proposals. This second approach incorporates the use of probabilistic lexicons as a source of translation information.

Blatz et al. (2003) introduce a more complex collection of features for word-level MTQE, using semantic features based on WordNet (Miller, 1995), translation probabilities from IBM model 1 (Brown et al., 1993), word posterior probabilities (Blatz et al., 2003), and alignment templates from statistical MT (SMT) models. All the features they use are combined to train a binary classifier which is used to determine the confidence scores.

Ueffing and Ney (2007) divide the features used by their approach in two types: those which are independent of the MT system used for translation (system-independent), and those which are extracted from internal data of the SMT system they use for translation (system-dependent). These features are obtained by comparing the output of an SMT system T_1 to a collection of alternative translations $\{T_i\}_{i=2}^{N_T}$ obtained by using the N-best list from the same SMT system. Several distance metrics are then used to check how often word t_j, the word in position j of T, is found in each translation alternative T_i, and how far from position j. These features rely on the assumption that a high occurrence frequency in a similar position is an evidence that t_j does not need to be post-edited. Biçici (2013) proposes a strategy for extending this kind of system-dependent features to what could be called a system-independent scenario. His approach consists in choosing parallel data from an additional parallel corpus which are close to the segment S to be translated by means of feature-decay algorithms (Biçici and Yuret, 2011). Once this parallel data are extracted, a new SMT system is trained and its internal data is used to extract these features.

The MULTILIZER approach to (sentence-level) MTQE (Bojar et al., 2014) also uses other MT systems to translate S into the TL and T into the SL. These translations are then used as a pseudo-reference and the similarity between them and the original SL and TL sentences is computed and taken as an indication of quality. This approach, as well as the one by Biçici and Yuret's (2011) are the most similar ones to our approach. One of the main differences is that they translate whole segments, whereas we translate sub-segments. As a result, we can obtain useful information about specific words in the translation. As the approach in this paper, MULTILIZER also combines several sources of bilingual information, while Biçici and Yuret (2011) only uses one MT system.[4]

Among the recent works on MTQE, it is worth mentioning the QuEst project (Specia et al., 2013), which sets a framework for MTQE, both at the sentence level and at the word level. This framework defines a large collection of features which can be divided in three groups: those measuring the complexity of the SL segment S, those measuring the confidence on the MT system, and those measuring both fluency and adequacy directly on the

[4] To the best of our knowledge, there is not any public description of the internal workings of MULTILIZIER.

translation hypothesis T. In fact, some of the most successful approaches in the word-level MTQE task in the Workshop on Statistical Machine Translation in 2014 (WMT 2014) (Bojar et al., 2014) are based on some of the features defined in that framework (Camargo de Souza et al., 2014).

The work described in this paper is aimed at being a system-independent approach that uses available on-line bilingual resources for word-level MTQE. This work is inspired by the work by Esplà-Gomis et al. (2011), in which several on-line MT systems are used for word-level quality estimation in translation-memory-based computer aided translation tasks. In the work by Esplà-Gomis et al. (2011), given a translation unit (S, T) suggested to the translator for the SL segment to be translated S', MT is used to translate sub-segments from S into the TL, and TL sub-segments from T into the SL. Sub-segment pairs obtained through MT that are found both in S and T are an evidence that they are related. The alignment between S and S', together with the sub-segment translations between S and T help to decide which words in T should be modified to get T', the desired translation of S'. Based on the same idea, we built a brand-new collection of word-level features to extend this approach to MTQE. One of the main advantages of this approach as compared to other approaches described in this section is that it uses light bilingual information extracted from any available source. Obtaining this information directly from the Internet allows us to obtain on the fly confidence estimates for the words in T without having to rely on more complex sources, such as probabilistic lexicons, part-of-speech information or word nets.

3 Word-level quality estimation using bilingual sources of information from the Internet

The approach proposed in this work for word-level MTQE uses binary classification based on features obtained through sources of bilingual information available on-line. We use these sources of bilingual information to detect connections between the original SL segment S and a given translation hypothesis T in the TL following the same method proposed by Esplà-Gomis et al. (2011): all the overlapping sub-segments of S and T, up to a given length L, are obtained and translated into the TL and the SL, respectively, using the sources of bilingual information available. The resulting collections of sub-segment translations $M_{S\rightarrow T}$ and $M_{T\rightarrow S}$ can be then used to spot sub-segment correspondences between T and S. In this section we describe a collection of features designed to identify these relations for their exploitation for word-level MTQE.

Positive features. Given a collection of sub-segment translations M (either $M_{S\rightarrow T}$ or $M_{T\rightarrow S}$), one of the most obvious features consists in computing the amount of sub-segment translations $(\sigma, \tau) \in M$ that confirm that word t_j in T should be kept in the translation of S. We consider that a sub-segment translation (σ, τ) confirms t_j if σ is a sub-segment of S, and τ is a sub-segment of T that covers position j. Based on this idea, we propose the collection of positive features Pos_n:

$$\mathrm{Pos}_n(j, S, T, M) = \frac{|\{\tau : \tau \in \mathrm{conf}_n(j, S, T, M)\}|}{|\{\tau : \tau \in \mathrm{seg}_n(T) \land j \in \mathrm{span}(\tau, T)\}|}$$

where $\mathrm{seg}_n(X)$ represents the set of all possible n-word sub-segments of segment X and function $\mathrm{span}(\tau, T)$ returns the set of word positions spanned by the sub-segment τ in the segment T.[5] Function $\mathrm{conf}_n(j, S, T, M)$ returns the collection of sub-segment pairs (σ, τ) that confirm a given word t_j, and is defined as:

$$\mathrm{conf}_n(j, S, T, M) = \{(\sigma, \tau) \in M : \tau \in \mathrm{seg}_n(T) \land \sigma \in \mathrm{seg}_*(S) \land j \in \mathrm{span}(\tau, T)\}$$

where $\mathrm{seg}_*(X)$ is similar to $\mathrm{seg}_n(X)$ but without length constraints.[6]

Additionally, we propose a second collection of features, which use the information about the translation frequency between the pairs of sub-segments in M. This information is not available for MT, although it is for the bilingual concordancer we have used (see Section 4). This frequency determines how often σ is translated as τ and, therefore, how reliable this translation is. We define $\mathrm{Pos}_n^{\mathrm{freq}}$ to obtain these features as:

$$\mathrm{Pos}_n^{\mathrm{freq}}(j, S, T, M) = \sum_{\forall(\sigma,\tau)\in \mathrm{conf}_n(j,S,T,M)} \frac{\mathrm{occ}(\sigma, \tau, M)}{\sum_{\forall(\sigma,\tau')\in M} \mathrm{occ}(\sigma, \tau', M)}$$

where function $\mathrm{occ}(\sigma, \tau, M)$ returns the number of occurrences in M of the sub-segment pair (σ, τ).

[5] Note that a sub-segment τ may be found more than once in segment T: function $\mathrm{span}(\tau, T)$ returns all the possible positions spanned.

[6] Two variants of function conf_n were tried: one applying also length constraints when segmenting S (with the consequent increment in the number of features), and one not applying length constraints at all. Preliminary results confirmed that constraining only the length of τ was the best choice.

Both positive features, $\mathrm{Pos}(\cdot)$ and $\mathrm{Pos}^{\mathrm{freq}}(\cdot)$, are computed for t_j for all the values of sub-segment length n up to L. In addition, they can be computed for both $M_{S\to T}$ and $M_{T\to S}$, producing $4L$ positive features in total for each word t_j.

Negative features. Our negative features, i.e. those features that help to identify words that should be post-edited in the translation hypothesis T, are also based on sub-segment translations $(\sigma,\tau) \in M$, but they are used in a different way. Negative features use those sub-segments τ that fit two criteria: (a) they are the translation of a sub-segment σ from S but cannot be matched in T; and (b) when they are aligned to T using the Levenshtein edit distance algorithm (Levenshtein, 1966), both their first word θ_1 and last word $\theta_{|\tau|}$ can be aligned, therefore delimiting a sub-segment τ' of T. Our hypothesis is that those words t_j in τ' which cannot be aligned to τ are likely to need to be post-edited. We define our negative feature collection $\mathrm{Neg}_{mn'}$ as:

$$\mathrm{Neg}_{mn'}(j,S,T,M) = \sum_{\forall \tau \in \mathrm{NegEvidence}_{mn'}(j,S,T,M)} \frac{1}{\mathrm{alignmentsize}(\tau,T)}$$

where $\mathrm{alignmentsize}(\tau,T)$ returns the length of the sub-segment τ' delimited by τ in T. Function $\mathrm{NegEvidence}_{mn'}(\cdot)$ returns the set of τ sub-segments that are considered negative evidence and is defined as:

$$\begin{aligned}\mathrm{NegEvidence}_{mn'}(j,S,T,M) = \{\tau : (\sigma,\tau) \in M \\ \wedge\, \sigma \in \mathrm{seg}_m(S) \,\wedge\, |\tau'| = n' \,\wedge \\ \tau \notin \mathrm{seg}_*(T) \,\wedge\, \mathrm{IsNeg}(j,\tau,T)\}\end{aligned}$$

In this function length constraints are set so that sub-segments σ take lengths $m \in [1,L]$.[7] However, the case of the sub-segments τ is slightly different: n' does not stand for the length of the sub-segments, but the number of words in τ which are aligned to T.[8] Function $\mathrm{IsNeg}(\cdot)$ defines the set of conditions required to consider a sub-segment τ a negative evidence for word t_j:

$$\begin{aligned}\mathrm{IsNeg}(j,\tau,T) = \exists j',j'' \in [1,|T|] : j' < j < j'' \\ \wedge\, \mathrm{aligned}(t_{j'},\theta_1) \,\wedge\, \mathrm{aligned}(t_{j''},\theta_{|\tau|}) \,\wedge \\ \nexists \theta_k \in \mathrm{seg}_1(\tau) : \mathrm{aligned}(t_j,\theta_k)\end{aligned}$$

where $\mathrm{aligned}(X,Y)$ is a binary function that checks whether words X and Y are aligned or not.

Negative features $\mathrm{Neg}_{mn'}(\cdot)$ are computed for t_j for all the values of SL sub-segment lengths $m \in [1,L]$ and the number of TL words $n' \in [2,L]$ which are aligned to words θ_k in sub-segment τ. Note that the number of aligned words between T and τ cannot be lower than 2 given the constraints set by function $\mathrm{IsNeg}(j,\tau,T)$. This results in a collection of $L \times (L-1)$ negative features. Obviously, for these features only $M_{S\to T}$ is used, since in $M_{T\to S}$ all the sub-segments τ can be found in T.

4 Experimental setting

The experiments described in this section compare the results of our approach to those in the word-level MTQE task in WMT 2014 (Bojar et al., 2014), which are considered the state of the art in the task. In this section we describe the sources of bilingual information used for our experiments, as well as the binary classifier and the data sets used for evaluation.

4.1 Evaluation data sets

Four data sets for different language pairs were published for the word-level MTQE task in WMT 2014: English–Spanish (EN–ES), Spanish–English (ES–EN), English–German (EN–DE), and German–English (DE–EN). The data sets contain the original SL segments, and their corresponding translation hypotheses tokenised at the level of words. Each word is tagged by hand using three levels of granularity:

- **binary**: words are classified only taking into account if they need to be post-edited (class *BAD*) or not (class *OK*);
- **level 1**: extension of the binary classification which differentiates between *accuracy* errors and *fluency* errors;
- **multi-class**: fine-grained classification of errors divided in 20 categories.

In this work we focus on the binary classification, which is the base for the other classification granularities.

Four evaluation metrics were defined for this task:

- The F_1 score weighted by the rate ρ_c of instances of a given class c in the data set:

$$F_1^w = \sum_{\forall c \in C} \rho_c \frac{2 p_c r_c}{p_c + r_c}$$

[7] In contrast to the positive features, preliminary results showed an improvement in the performance of the classifier when constraining the length of the σ sub-segments used for each feature in the set.

[8] That is, the length of longest common sub-segment of τ and T.

where C is the collection of classes defined for a given level of granularity (OK and BAD for the binary classification) and p_c and r_c are the precision and recall for a class $c \in C$, respectively;

- The F_1 score of the less frequent class in the data set (class BAD, in the case of binary classification):

$$F_1^{\mathrm{BAD}} = \frac{2 \times p_{\mathrm{BAD}} \times r_{\mathrm{BAD}}}{p_{\mathrm{BAD}} + r_{\mathrm{BAD}}};$$

- The Matthews correlation coefficient (MCC), which takes values in $[-1, 1]$ and is more reliable than the F_1 score for unbalanced data sets (Powers, 2011):

$$\mathrm{MCC} = \frac{T_{\mathrm{OK}} \times T_{\mathrm{BAD}} - F_{\mathrm{OK}} \times F_{\mathrm{BAD}}}{\sqrt[2]{A_{\mathrm{OK}} \times A_{\mathrm{BAD}} \times P_{\mathrm{OK}} \times P_{\mathrm{BAD}}}}$$

  where T_{OK} and T_{BAD} stand for the number of instances correctly classified for each class, F_{OK} and F_{BAD} stand for the number of instances wrongly classified for each class, P_{OK} and P_{BAD} stand for the number of instances classified either as OK or BAD, and A_{OK} and A_{BAD} stand for the actual number of each class; and

- Total accuracy (ACC):

$$\mathrm{ACC} = \frac{T_{\mathrm{OK}} + T_{\mathrm{BAD}}}{P_{\mathrm{OK}} + P_{\mathrm{BAD}}}$$

The comparison between the approach presented in this work and those described by Bojar et al. (2014) is based on the F_1^{BAD} score because this was the main metric used to compare the different approaches participating in WMT 2014. However, all the metrics are reported for a better analysis of the results obtained.

4.2 Sources of Bilingual Information

As already mentioned, two different sources of information were used in this work, MT and a bilingual concordancer. For our experiments we used two MT systems which are freely available on the Internet: Apertium and Google Translate. These MT systems were exploited by translating the sub-segments, for each data set, in both directions (from SL to TL and vice versa). It is worth noting that language pairs EN–DE and DE–EN are not available for Apertium. For these data sets only Google Translate was used.

The bilingual concordancer *Reverso Context* was also used for translating sub-segments. Namely, the sub-sentential translation memory of this system was used, which is a much richer source of bilingual information and provides, for a given SL sub-segment, the collection of TL translation alternatives, together with the number of occurrences of the sub-segments pair in the translation memory. Furthermore, the sub-segment translations obtained from this source of information are more reliable, since they are extracted from manually translated texts. On the other hand, its main weakness is the coverage: although Reverso Context uses a large translation memory, no translation can be obtained for those SL sub-segments which cannot be found in it. In addition, the sub-sentential translation memory contains only those sub-segment translations with a minimum number of occurrences. On the contrary, MT systems will always produce a translation, even though it may be wrong or contain untranslated out-of-vocabulary words. Our hypothesis is that combining both sources of bilingual information can lead to reasonable results for word-level MTQE.

For our experiments, we computed the features described in Section 3 separately for both sources of information. The value of the maximum sub-segment length L used was set to 5, which resulted in a collection of 40 features from the bilingual concordancer, and 30 from MT.[9]

4.3 Binary classifier

Esplà-Gomis et al. (2011) use a simple perceptron classifier for word-level quality estimation in translation-memory-based computer-aided translation. In this work, a more complex *multilayer perceptron* (Duda et al., 2000, Section 6) is used, as implemented in Weka 3.6 (Hall et al., 2009). Multilayer perceptrons (also known as *feedforward neural networks*) have a complex structure which incorporates one or more *hidden layers*, consisting of a collection of perceptrons, placed between the input of the classifier (the features) and the output perceptron. This hidden layer makes multilayer perceptrons suitable for non-linear classification problems (Duda et al., 2000, Section 6). In fact, Hornik et al. (1989) proved that neural networks with a single hidden layer containing a finite number of neurons are universal approximators and may therefore be able to perform better than a simple per-

[9] As already mentioned, the features based on translation frequency cannot be obtained for MT.

ceptron for complex problems. In our experiments, we have used a batch training strategy, which iteratively updates the weights of each perceptron in order to minimise a total error function. A subset of 10% of the training examples was extracted from the training set before starting the training process and used as a validation set. The weights were iteratively updated on the basis of the error computed in the other 90%, but the decision to stop the training (usually referred as the convergence condition) was based on this validation set. This is a usual practice whose objective is to minimise the risk of overfitting. The training process stops when the total error obtained in an iteration is worse than that obtained in the previous 20 iterations.[10]

Hyperparameter optimisation was carried out using a grid search (Bergstra et al., 2011) in a 10-fold cross-validation fashion in order to choose the hyperparameters optimising the results for the metric to be used for comparison, F_1 for class *BAD*:

- *Number of nodes in the hidden layer*: Weka (Hall et al., 2009) makes it possible to choose from among a collection of predefined network designs; the design performing best in most cases happened to have the same number of nodes in the hidden layer as the number of features.

- *Learning rate*: this parameter allows the dimension of the weight updates to be regulated by applying a factor to the error function after each iteration; the value that best performed for most of our training data sets was 0.9.

- *Momentum*: when updating the weights at the end of a training iteration, momentum smooths the training process for faster convergence by making it dependent on the previous weight value; in the case of our experiments, it was set to 0.07.

5 Results and discussion

language pair	weighted F_1	BAD F_1	MCC	accuracy
EN–ES	18.71	52.53	0.00	35.62
ES–EN	5.28	29.98	0.00	17.63
EN–ES	12.78	44.57	0.00	28.67
DE–EN	8.20	36.60	0.00	22.40

Table 1: Results of the "always BAD" baseline for the different data sets.

language pair	weighted F_1	BAD F_1	MCC	accuracy
EN–ES	62.00	48.73	18.23	61.62
ES–EN	79.54	29.14	25.47	82.98
EN–DE	71.51	45.30	28.61	72.97
DE–EN	72.41	26.13	16.08	76.14

Table 2: Results of the best performing systems for the different data sets according to the results published by Bojar et al. (2014).

Table 1 shows the results obtained by the baseline consisting on marking all the words as BAD, whereas Table 2 shows the reference results obtained by the best performing system according to the results published by Bojar et al. (2014). These tables are used as a reference for the results obtained with the approach described in this work.

Table 3 shows the results obtained when using Reverso Context as the only source of information. Using only Reverso Context leads to reasonably good results for language pairs EN–ES and EN–DE, while for the other two language pairs results are much worse, basically because no word was classified as needing to be post-edited. This situation is caused by the fact that, in both cases, the amount of examples of words to be post-edited in the training set is very small (lower than 21%). In this case, if the features are not informative enough, the strong bias leads to a classifier that always recommends to keep all words untouched. However, it is worth noting that with a small amount of features (40 features) state-of-the-art results were obtained for two data sets.[11] Namely, in the case of the EN–ES data set, the one with the largest amount of training instances, the results for the main metric (F_1 score for the less frequent class, in this case *BAD*) were better than those of the state of the art. In the case of the EN–DE data set the results are noticeably lower than the state of the art, but they are still comparable to them.

Table 4 shows the results obtained when combining the information from Reverso Context and the MT systems Apertium and Google Translate. Again, one of the best results is obtained for the EN–ES data set, which would again beat the state of the art for the F_1 score for the *BAD* class, and

[10] It is usual to set a number of additional iterations after the error stops improving, in case the function is in a local minimum, and the error starts decreasing again after a few more iterations. If the error continues to worsen after these 20 iterations, the weights used are those obtained after the iteration with the lowest error.

[11] We focus our comparison on the F_1 score for the *BAD* class because this was the metric on which the classifiers were optimised.

language pair	weighted F_1	BAD F_1	MCC	accuracy
EN–ES	60.18	49.09	16.28	59.46
ES–EN	74.41	0.00	0.00	82.37
EN–DE	65.88	41.24	17.05	65.71
DE–EN	67.82	0.00	0.00	77.60

Table 3: Results of the approach proposed in this paper for the same data sets used to obtain Table 2 using Reverso Context as the only source of bilingual information.

language pair	weighted F_1	BAD F_1	MCC	accuracy
EN–ES	61.43	49.03	17.71	60.91
ES–EN	75.87	10.44	9.61	81.82
EN–DE	66.75	43.07	19.38	78.71
DE–EN	75.00	40.33	25.85	76.03

Table 4: Results of the approach proposed in this work for the same data sets used to obtain Table 2 using both Reverso Context and both Google Translate and Apertium as the sources of bilingual information.

which obtained results still closer to those of the state of the art for the rest of metrics. In addition, the biased classification problem for data sets DE–EN and ES–EN is alleviated. Actually, the results for the DE–EN language pair are particularly good, and outperform the state of the art for all the metrics. The low F_1 score obtained for the ES–EN data set may be explained by the unbalanced amount of positive and negative instances. Actually, the ratio of negative instances is somewhat related to the results obtained: 35% for EN–ES, 17% for ES–EN, 30% for EN–DE and 21% for DE–EN. A closer analysis of the results shows that our approach is better when detecting errors in the *Terminology*, *Mistranslation*, and *Unintelligible* subclasses. The ratio of this kind of errors over the total amount of negative instances for each data set is again related to the results obtained: 73% for EN–ES, 27% for ES–EN, 47% for EN–DE and 35% for DE–EN. This information may explain the differences in the results obtained for each data set.

Again, it is worth noting that this light method using a reduced set of 70 features can obtain, for most of the data sets, results comparable to those obtained by approaches using much more features. For example, the best system for the data set EN–ES (Camargo de Souza et al., 2014) used 163 features, while the winner system for the rest of data sets (Biçici and Way, 2014; Biçici, 2013) used 511,000 features. The sources of bilingual information used in this work are rather rich; however, given that any source of bilingual information could be used on the fly, simpler sources of bilingual information could also be used. It would therefore be interesting to carry out a deeper evaluation of the impact of the type and quality of the resources used with this approach.

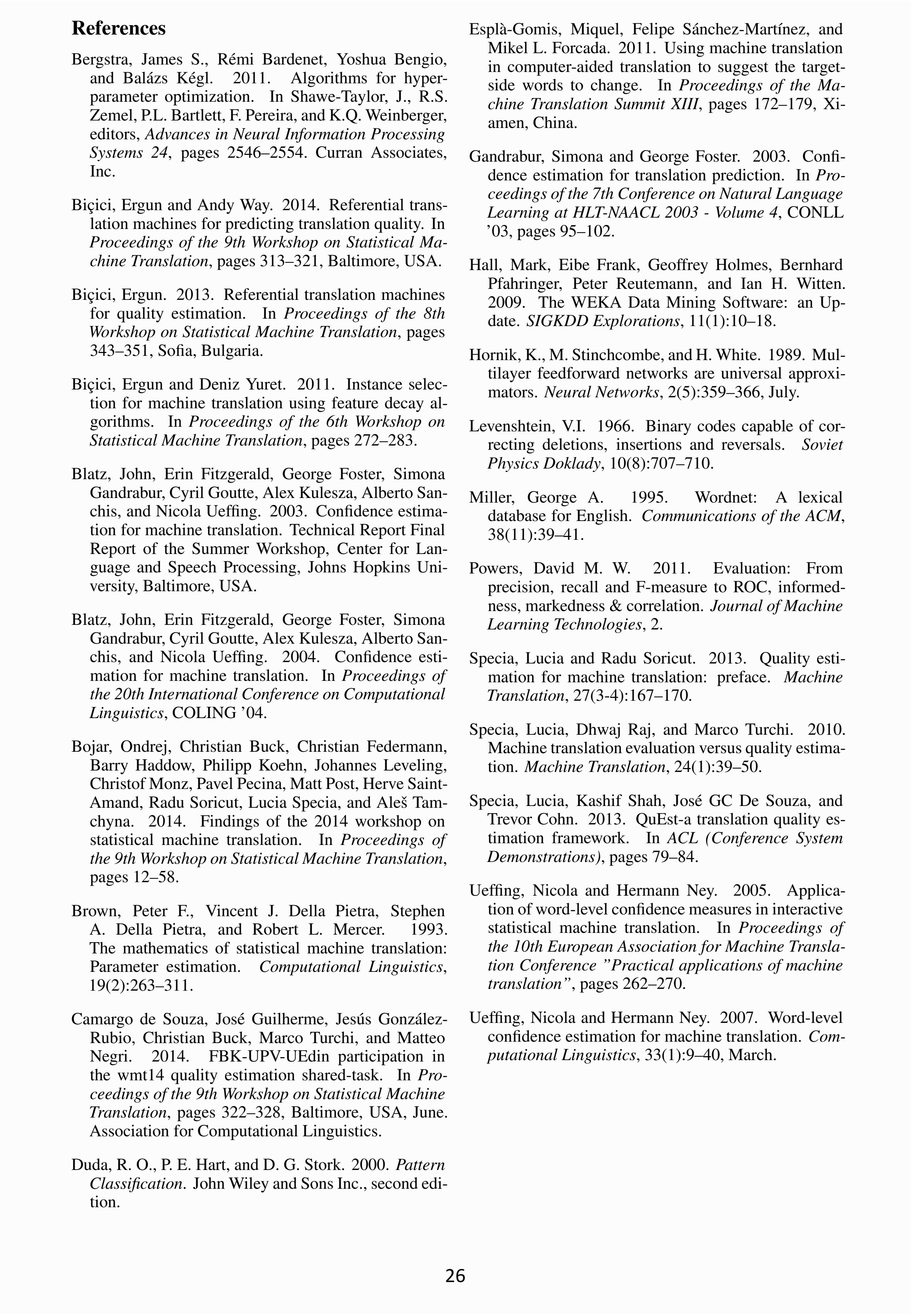
6 Concluding remarks

In this paper we describe a novel approach for word-level MTQE based on the use of on-line available bilingual resources. This approach is aimed at being system-independent, since it does not make any assumptions about the MT system used for producing the translation hypotheses to be evaluated. Furthermore, given that this approach can use any source of bilingual information as a black box, it can be easily used with few resources. In addition, adding new sources of information is straightforward, providing considerable room for improvement. The results described in Section 5 confirm that our approach can reach results comparable to those in the state of the art using a smaller collection of features than those used by most of the other approaches.

Although the results described in this paper are encouraging, it is worth noting that it is difficult to extract strong conclusions from the small data sets used. A wider evaluation should be done, involving larger data sets and more language pairs. As future work, we plan to extend this method by using other on-line resources to improve the on-line coverage when spotting sub-segment translations; namely, different bilingual concordancers and on-line dictionaries. Monolingual target-language information could also be obtained from the Internet to deal with fluency issues, for example, getting the frequency of a given n-gram from search engines. We will also study the combination of these features with features used in previous state-of-the-art systems (see Section 2) Finally, it would be interesting to try the new features defined here in word-level quality estimation for computer-aided translation tools, as in Esplà-Gomis et al. (2011).

Acknowledgements

Work partially funded by the Spanish Ministerio de Ciencia e Innovación through projects TIN2009-14009-C02-01 and TIN2012-32615 and by the European Commission through project PIAP-GA-2012-324414 (Abu-MaTran). We specially thank Reverso-Softissimo and Prompsit Language Engineering for providing the access to Reverso Context, and to the University Research Program for Google Translate that granted us access to the Google Translate service.



References

Bergstra, James S., Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for hyper-parameter optimization. In Shawe-Taylor, J., R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2546–2554. Curran Associates, Inc.

Biçici, Ergun and Andy Way. 2014. Referential translation machines for predicting translation quality. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pages 313–321, Baltimore, USA.

Biçici, Ergun. 2013. Referential translation machines for quality estimation. In *Proceedings of the 8th Workshop on Statistical Machine Translation*, pages 343–351, Sofia, Bulgaria.

Biçici, Ergun and Deniz Yuret. 2011. Instance selection for machine translation using feature decay algorithms. In *Proceedings of the 6th Workshop on Statistical Machine Translation*, pages 272–283.

Blatz, John, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2003. Confidence estimation for machine translation. Technical Report Final Report of the Summer Workshop, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA.

Blatz, John, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04.

Bojar, Ondrej, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pages 12–58.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Camargo de Souza, José Guilherme, Jesús González-Rubio, Christian Buck, Marco Turchi, and Matteo Negri. 2014. FBK-UPV-UEdin participation in the wmt14 quality estimation shared-task. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pages 322–328, Baltimore, USA, June. Association for Computational Linguistics.

Duda, R. O., P. E. Hart, and D. G. Stork. 2000. *Pattern Classification*. John Wiley and Sons Inc., second edition.

Esplà-Gomis, Miquel, Felipe Sánchez-Martínez, and Mikel L. Forcada. 2011. Using machine translation in computer-aided translation to suggest the target-side words to change. In *Proceedings of the Machine Translation Summit XIII*, pages 172–179, Xiamen, China.

Gandrabur, Simona and George Foster. 2003. Confidence estimation for translation prediction. In *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 95–102.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: an Update. *SIGKDD Explorations*, 11(1):10–18.

Hornik, K., M. Stinchcombe, and H. White. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, July.

Levenshtein, V.I. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Miller, George A. 1995. Wordnet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Powers, David M. W. 2011. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2.

Specia, Lucia and Radu Soricut. 2013. Quality estimation for machine translation: preface. *Machine Translation*, 27(3-4):167–170.

Specia, Lucia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50.

Specia, Lucia, Kashif Shah, José GC De Souza, and Trevor Cohn. 2013. QuEst-a translation quality estimation framework. In *ACL (Conference System Demonstrations)*, pages 79–84.

Ueffing, Nicola and Hermann Ney. 2005. Application of word-level confidence measures in interactive statistical machine translation. In *Proceedings of the 10th European Association for Machine Translation Conference "Practical applications of machine translation"*, pages 262–270.

Ueffing, Nicola and Hermann Ney. 2007. Word-level confidence estimation for machine translation. *Computational Linguistics*, 33(1):9–40, March.

UAlacant word-level machine translation quality estimation system at WMT 2015

Miquel Esplà-Gomis Felipe Sánchez-Martínez Mikel L. Forcada
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain
{mespla,fsanchez,mlf}@dlsi.ua.es

Abstract

This paper describes the Universitat d'Alacant submissions (labelled as UAlacant) for the machine translation quality estimation (MTQE) shared task in WMT 2015, where we participated in the word-level MTQE sub-task. The method we used to produce our submissions uses external sources of bilingual information as a *black box* to spot sub-segment correspondences between a source segment S and the translation hypothesis T produced by a machine translation system. This is done by segmenting both S and T into overlapping sub-segments of variable length and translating them in both translation directions, using the available sources of bilingual information *on the fly*. For our submissions, two sources of bilingual information were used: machine translation (Apertium and Google Translate) and the bilingual concordancer Reverso Context. After obtaining the sub-segment correspondences, a collection of features is extracted from them, which are then used by a binary classifer to obtain the final "GOOD" or "BAD" word-level quality labels. We prepared two submissions for this year's edition of WMT 2015: one using the features produced by our system, and one combining them with the baseline features published by the organisers of the task, which were ranked third and first for the sub-task, respectively.

1 Introduction

Machine translation (MT) post-editing is nowadays an indispensable step that allows to use machine translation for dissemination. Consequently, MT quality estimation (MTQE) (Blatz et al., 2004; Specia et al., 2010; Specia and Soricut, 2013) has emerged as a mean to minimise the post-editing effort by developing techniques that allow to estimate the quality of the translation hypotheses produced by an MT system. In order to boost the scientific efforts on this problem, the WMT 2015 MTQE shared task proposes three tasks that allow to compare different approaches at three different levels: segment-level (sub-task 1), word-level (sub-task 2), and document-level (sub-task 3).

Our submissions tackle the word-level MTQE sub-task, which proposes a framework for evaluating and comparing different approaches. This year, the sub-task used a dataset obtained by translating segments in English into Spanish using MT. The task consists in identifying which words in the translation hypothesis had to be post-edited and which of them had to be kept unedited by applying the labels "BAD" and "GOOD", respectively. In this paper we describe the approach behind the two submissions of the Universitat d'Alacant team to this sub-task. For our submissions we applied the approach proposed by Esplà-Gomis et al. (2015b), who use black-box bilingual resources from the Internet for word-level MTQE. In particular, we combined two on-line MT systems, Apertium[1] and Google Translate,[2] and the bilingual concordancer Reverso Context[3] to spot sub-segment correspondences between a sentence S in the source language (SL) and a given translation hypothesis T in the target language (TL). To do so, both S and T are segmented into all possible overlapping sub-

[1] http://www.apertium.org

[2] http://translate.google.com

[3] http://context.reverso.net/translation/

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 309–315,
Lisboa, Portugal, 17-18 September 2015. ©2015 Association for Computational Linguistics.

segments up to a certain length and translated into the TL and the SL, respectively, by means of the sources of bilingual information mentioned above. These sub-segment correspondences are used to extract a collection of features that is then used by a binary classifier to determine the final word-level MTQE labels.

One of the novelties of the task this year is that the organisation provided a collection of baseline features for the dataset published. Therefore, we submitted two systems: one using only the features defined by Esplà-Gomis et al. (2015b), and another combining them with the baseline features published by the organisers of the shared task. The results obtained by our submissions were ranked third and first, respectively.

The rest of the paper is organised as follows. Section 2 describes the approach used to produce our submissions. Section 3 describes the experimental setting and the results obtained. The paper ends with some concluding remarks.

2 Sources of bilingual information for word-level MTQE

The approach proposed by Esplà-Gomis et al. (2015b), which is the one we have followed in our submissions for the MTQE shared task in WMT 2015, uses binary classification based on a collection of features computed for each word by using available sources of bilingual information. These sources of bilingual information are obtained from on-line tools and are used on-the-fly to detect relations between the original SL segment S and a given translation hypothesis T in the TL. This method has been previously used by the authors in other cross-lingual NLP tasks, such as word-keeping recommendation (Esplà-Gomis et al., 2015a) or cross-lingual textual entailment (Esplà-Gomis et al., 2012), and consists of the following steps: first, all the overlapping sub-segments σ of S up to given length L are obtained and translated into the TL using the sources of bilingual information available. The same process is carried out for all the overlapping sub-segments τ of T, which are translated into the SL. The resulting collections of sub-segment translations $M_{S\to T}$ and $M_{T\to S}$ are then used to spot sub-segment correspondences between T and S. In this section we describe a collection of features designed to identify these relations for their exploitation for word-level MTQE.

2.1 Positive features

Given a collection of sub-segment translations $M = \{\sigma, \tau\}$, such as the collections $M_{S\to T}$ and $M_{T\to S}$) described above, one of the most obvious features consists in computing the amount of sub-segment translations $(\sigma, \tau) \in M$ that confirm that word t_j in T should be kept in the translation of S. We consider that a sub-segment translation (σ, τ) confirms t_j if σ is a sub-segment of S, and τ is a sub-segment of T that covers position j. Based on this idea, we propose the collection of positive features Pos_n:

$$\text{Pos}_n(j, S, T, M) = \frac{|\{\tau : (\sigma, \tau) \in \text{conf}_n(j, S, T, M)\}|}{|\{\tau : \tau \in \text{seg}_n(T) \land j \in \text{span}(\tau, T)\}|}$$

where $\text{seg}_n(X)$ represents the set of all possible n-word sub-segments of segment X and function $\text{span}(\tau, T)$ returns the set of word positions spanned by the sub-segment τ in the segment T.[4] Function $\text{conf}_n(j, S, T, M)$ returns the collection of sub-segment pairs (σ, τ) that confirm a given word t_j, and is defined as:

$$\text{conf}_n(j, S, T, M) = \{(\sigma, \tau) \in M : \tau \in \text{seg}_n(T) \land \sigma \in \text{seg}_*(S) \land j \in \text{span}(\tau, T)\}$$

where $\text{seg}_*(X)$ is similar to $\text{seg}_n(X)$ but without length constraints.[5]

We illustrate this collection of features with an example. Suppose the Catalan segment S =“Associació Europea per a la Traducció Automàtica”, an English translation hypothesis T =“European Association for the Automatic Translation”, and the most adequate (reference) translation T'=“European Association for Machine Translation”. According to the reference, the words *the* and *Automatic* in the translation hypothesis should be marked as BAD: *the* should be removed and *Automatic* should be replaced by *Machine*. Finally, suppose that the collection $M_{S\to T}$ of sub-segment pairs (σ, τ) is obtained by applying the available sources of bilingual information to translate into English the sub-segments in S up to length 3:[6]

[4] Note that a sub-segment τ may be found more than once in segment T: function $\text{span}(\tau, T)$ returns all the possible positions spanned.

[5] Esplà-Gomis et al. (2015b) conclude that constraining only the length of τ leads to better results than constraining both σ and τ.

[6] The other translation direction is omitted for simplicity.

$M_{S\to T} = \{$***(“Associació”, “Association”),*** ***(“Europea”, “European”),*** ***(“per”, “for”),*** *(“a”, “to”),* ***(“la”, “the”),*** ***(“Traducció”, “Translation”),*** ***(“Automàtica”, “Automatic”),*** ***(“Associació Europea”, “European Association”),*** *(“Europea per”, “European for”),* ***(“per a”, “for”),*** *(“a la”, “to the”),* *(“la Traducció”, “the Translation”),* *(“Traducció Automàtica”, “Machine Translation”),* ***(“Associació Europea per”, “European Association for”),*** *(“Europea per a”, “European for the”),* ***(“per a la”, “ for the”),*** *(“a la Traducció”, “to the Translation”),* *(“la Traducció Automàtica”, “the Machine Translation”)*$\}$

Note that the sub-segment pairs (σ, τ) in bold are those confirming the translation hypothesis T, while the rest contradict some parts of the hypothesis. For the word *Machine* (which corresponds to word position 5), there is only one sub-segment pair confirming it *(“Automàtica”, “Automatic”)* with length 1, and no one with lengths 2 and 3. Therefore, we have that:

$$\mathrm{conf}_1(5, S, T, M) = \{(\textit{“Automàtica”}, \textit{“Automatic”})\}$$

$$\mathrm{conf}_2(5, S, T, M) = \emptyset$$

$$\mathrm{conf}_3(5, S, T, M) = \emptyset$$

In addition, we have that the sub-segments τ in $\mathrm{seg}_*(T)$ covering the word *Automatic* for lengths in $[1, 3]$ are:

$$\{\tau : \tau \in \mathrm{seg}_1(T) \wedge j \in \mathrm{span}(\tau, T)\} = \{\textit{“Automatic”}\}$$

$$\{\tau : \tau \in \mathrm{seg}_2(T) \wedge j \in \mathrm{span}(\tau, T)\} = \{\textit{“the Automatic”}, \textit{“Automatic Translation”}\}$$

$$\{\tau : \tau \in \mathrm{seg}_3(T) \wedge j \in \mathrm{span}(\tau, T)\} = \{\textit{“for the Automatic”}, \textit{“the Automatic Translation”}\}$$

Therefore, the resulting positive features for this word would be:

$$\mathrm{Pos}_1(5, S, T, M) = \frac{\mathrm{conf}_3(5, S, T, M)}{\{\tau : \tau \in \mathrm{seg}_1(T) \wedge j \in \mathrm{span}(\tau, T)\}} = \frac{1}{1}$$

$$\mathrm{Pos}_2(5, S, T, M) = \frac{\mathrm{conf}_2(5, S, T, M)}{\{\tau : \tau \in \mathrm{seg}_2(T) \wedge j \in \mathrm{span}(\tau, T)\}} = \frac{0}{2}$$

$$\mathrm{Pos}_3(5, S, T, M) = \frac{\mathrm{conf}_3(5, S, T, M)}{\{\tau : \tau \in \mathrm{seg}_3(T) \wedge j \in \mathrm{span}(\tau, T)\}} = \frac{0}{2}$$

A second collection of features, which use the information about the translation frequency between the pairs of sub-segments in M is also used. This information is not available for MT, but it is for the bilingual concordancer we have used (see Section 3). This frequency determines how often σ is translated as τ and, therefore, how reliable this translation is. We define $\mathrm{Pos}_n^{\mathrm{freq}}$ to obtain these features as:

$$\mathrm{Pos}_n^{\mathrm{freq}}(j, S, T, M) = \sum_{\forall(\sigma,\tau)\in \mathrm{conf}_n(j,S,T,M)} \frac{\mathrm{occ}(\sigma, \tau, M)}{\sum_{\forall(\sigma,\tau')\in M} \mathrm{occ}(\sigma, \tau', M)}$$

where function $\mathrm{occ}(\sigma, \tau, M)$ returns the number of occurrences in M of the sub-segment pair (σ, τ).

Following the running example, we may have an alternative and richer source of bilingual information, such as a sub-segmental translation memory, which contains 99 occurrences of word *Automàtica* translated as *Automatic*, as well as the following alternative translations: *Machine* (11 times), and *Mechanic* (10 times). Therefore, the positive feature using these frequencies for sub-segments of length 1 would be:

$$\mathrm{Pos}_1^{\mathrm{freq}}(5, S, T, M) = \frac{99}{99 + 11 + 10} = 0.825$$

Both positive features, $\mathrm{Pos}(\cdot)$ and $\mathrm{Pos}^{\mathrm{freq}}(\cdot)$, are computed for t_j for all the values of sub-segment length $n \in [1, L]$. In addition, they can be computed for both $M_{S\to T}$ and $M_{T\to S}$; this yields $4L$ positive features in total for each word t_j.

2.2 Negative features

The negative features, i.e. those features that help to identify words that should be post-edited in the translation hypothesis T, are also based on sub-segment translations $(\sigma, \tau) \in M$, but they are used in a different way. Negative features use those sub-segments τ that fit two criteria: (a) they are the translation of a sub-segment σ from S but are not sub-segments of T; and (b) when they are aligned to T using the edit-distance algorithm (Wagner and Fischer, 1974), both their first word θ_1 and last word $\theta_{|\tau|}$ can be aligned, therefore delimiting a sub-segment τ' of T. Our hypothesis is that those words t_j in τ' which cannot be aligned to τ are likely to need postediting. We define our negative feature collection $\text{Neg}_{mn'}$ as:

$$\text{Neg}_{mn'}(j, S, T, M) = \sum_{\forall \tau \in \text{NegEvidence}_{mn'}(j,S,T,M)} \frac{1}{\text{alignmentsize}(\tau, T)}$$

where $\text{alignmentsize}(\tau, T)$ returns the length of the sub-segment τ' delimited by τ in T. Function $\text{NegEvidence}_{mn'}(\cdot)$ returns the set of sub-segments τ of T that are considered negative evidence and is defined as:

$$\begin{aligned}\text{NegEvidence}_{mn'}(j, S, T, M) = \{\tau : (\sigma, \tau) \in M \\ \wedge\, \sigma \in \text{seg}_m(S) \,\wedge\, |\tau'| = n' \,\wedge \\ \tau \notin \text{seg}_*(T) \,\wedge\, \text{IsNeg}(j, \tau, T)\}\end{aligned}$$

In this function length constraints are set so that sub-segments σ take lengths $m \in [1, L]$. While for the positive features, only the length of τ was constrained, the experiments carried out by Esplà-Gomis et al. (2015b) indicate that for the negative features, it is better to constrain also the length of σ. On the other hand, the case of the sub-segments τ is slightly different: n' does not stand for the length of the sub-segments, but the number of words in τ which are aligned to T.[7] Function $\text{IsNeg}(\cdot)$ defines the set of conditions required to consider a sub-segment τ a negative evidence for word t_j:

$$\begin{aligned}\text{IsNeg}(j, \tau, T) = \exists j', j'' \in [1, |T|] : j' < j < j'' \\ \wedge\, \text{aligned}(t_{j'}, \theta_1) \,\wedge\, \text{aligned}(t_{j''}, \theta_{|\tau|}) \,\wedge \\ \nexists \theta_k \in \text{seg}_1(\tau) : \text{aligned}(t_j, \theta_k)\end{aligned}$$

where $\text{aligned}(X, Y)$ is a binary function that checks whether words X and Y are aligned or not.

For our running example, only two sub-segment pairs (σ, τ) fit the conditions set by function $\text{IsNeg}(j, \tau, T)$ for the word *Automatic*: *("la Traducció", "the Translation")*, and *("la Traducció Automàtica", "the Machine Translation")*. As can be seen, for both (σ, τ) pairs, the words *the* and *Translation* in the sub-segments τ can be aligned to the words in positions 4 and 6 in T, respectively, which makes the number of words aligned $n' = 2$. In this way, we would have the evidences:

$$\text{NegEvidence}_{2,2}(5, S, T, M) = \{\text{“the Translation”}\}$$

$$\text{NegEvidence}_{3,2}(5, S, T, M) = \{\text{“the Machine Translation”}\}$$

As can be seen, in the case of sub-segment τ = "the Translation", these alignments suggest that word *Automatic* should be removed, while for the sub-segment τ = the Machine Translation" they suggest that word *Automatic* should be replaced by word *Machine*. The resulting negative features are:

$$\text{Neg}_{2,2}(5, S, T, M) = \tfrac{1}{3}$$

$$\text{Neg}_{3,2}(5, S, T, M) = \tfrac{1}{3}$$

Negative features $\text{Neg}_{mn'}(\cdot)$ are computed for t_j for all the values of SL sub-segment lengths $m \in [1, L]$ and the number of TL words $n' \in [2, L]$ which are aligned to words θ_k in sub-segment τ. Note that the number of aligned words between T and τ cannot be smaller than 2 given the constraints set by function $\text{IsNeg}(j, \tau, T)$. This results in a collection of $L \times (L - 1)$ negative features. Obviously, for these features only $M_{S \to T}$ is used, since in $M_{T \to S}$ all the sub-segments τ can be found in T.

3 Experiments

This section describes the dataset provided for the word-level MTQE sub-task and the results obtained by our method on these datasest. This year, the task consisted in measuring the word-level MTQE on a collection of segments in Spanish that had been obtained through machine translation from English. The organisers provided a dataset consisting of:

- *training set*: a collection of 11,272 segments in English (S) and their corresponding machine translations in Spanish (T); for every word in T, a label was provided: BAD for the words to be post-edited, and GOOD for those to be kept unedited;
- *development set*: 1,000 pairs of segments (S, T) with the corresponding MTQE labels that can be used to optimise the binary classifier trained by using the training set;
- *test set*: 1,817 pairs of segments (S, T) for which the MTQE labels have to be estimated with the binary classifier trained on the training and the development sets.

[7] That is, the length of longest common sub-segment of τ and T.

3.1 Binary classifier

A *multilayer perceptron* (Duda et al., 2000, Section 6) was used for classification, as implemented in Weka 3.6 (Hall et al., 2009), following the approach by Esplà-Gomis et al. (2015b). A subset of 10% of the training examples was extracted from the training set before starting the training process and used as a validation set. The weights were iteratively updated on the basis of the error computed in the other 90%, but the decision to stop the training (usually referred as the convergence condition) was based on this validation set, in order to minimise the risk of overfitting. The error function used was based on the the optimisation of the metric used for ranking, i.e. the F_1^{BAD} metric.

Hyperparameter optimisation was carried out on the development set, by using a grid search (Bergstra et al., 2011) in order to choose the hyperparameters optimising the results for the metric to be used for comparison, F_1 for class *BAD*:

- *Number of nodes in the hidden layer*: Weka (Hall et al., 2009) makes it possible to choose from among a collection of predefined network designs; the design performing best in most cases happened to have a single hidden layer containing the same number of nodes in the hidden layer as the number of features.
- *Learning rate*: this parameter allows the dimension of the weight updates to be regulated by applying a factor to the error function after each iteration; the value that best performed for most of our training data sets was 0.1.
- *Momentum*: when updating the weights at the end of a training iteration, momentum smooths the training process for faster convergence by making it dependent on the previous weight value; in the case of our experiments, it was set to 0.03.

3.2 Evaluation

As already mentioned, two configurations of our system were submitted: one using only the features defined in Section 2, and one combining them with the baseline features. In order to obtain our features we used two sources of bilingual information, as already mentioned: MT and a bilingual concordancer. As explained above, for our experiments we used two MT systems which are freely available on the Internet: Apertium and Google Translate. The bilingual concordancer *Reverso Context* was also used for translating sub-segments. Actually, only the sub-sentential translation memory of this system was used, which provides the collection of TL translation alternatives for a given SL sub-segment, together with the number of occurrences of the sub-segments pair in the translation memory.

Four evaluation metrics were proposed for this task:

- The precision P^c, i.e. the fraction of instances correctly labelled among all the instances labelled as c, where c is the class assigned (either GOOD or BAD in our case);
- The recall R^c, i.e. the fraction of instances correctly labelled as c among all the instances that should be labelled as c in the test set;
- The F_1^c score, which is defined as

  $$F_1^c = \frac{2 \times P^c \times R^c}{P^c + R^c};$$

  although the F_1^c score is computed both for GOOD and for BAD, it is worth noting that the F_1 score for the less frequent class in the data set (label BAD, in this case) is used as the main comparison metric;
- The F_1^w score, which is the version of F_1^c weighted by the proportion of instances of a given class c in the data set:

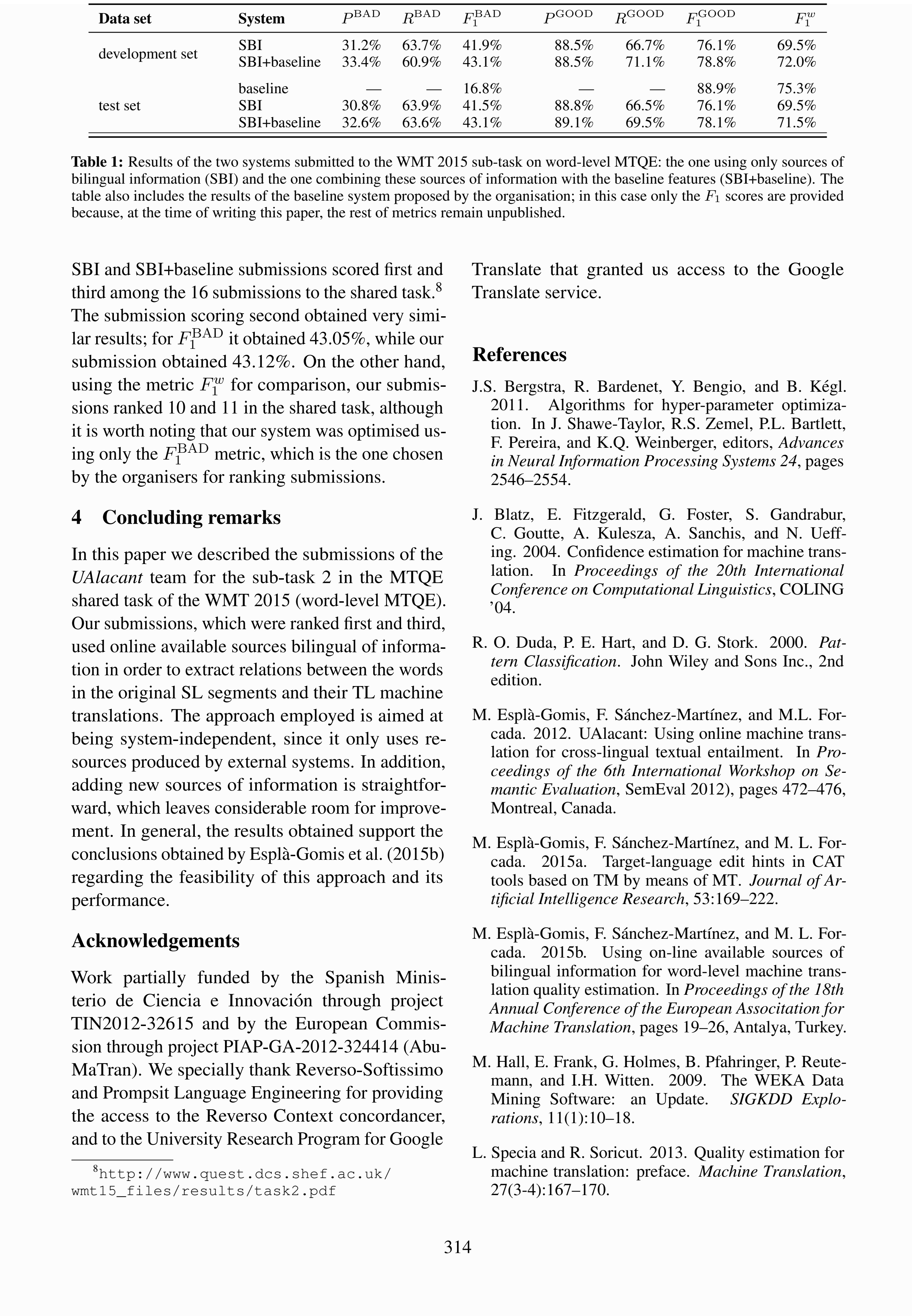
  $$F_1^w = \frac{N^{\text{BAD}}}{N^{\text{TOTAL}}} F_1^{\text{BAD}} + \frac{N^{\text{GOOD}}}{N^{\text{TOTAL}}} F_1^{\text{GOOD}}$$

  where N^{BAD} is the number of instances of the class BAD, N^{GOOD} is the number of instances of the class GOOD, and N^{TOTAL} is the total number of instances in the test set.

3.3 Results

Table 1 shows the results obtained by our system, both on the development set during the training phase and on the test set. The table also includes the results for the baseline system as published by the organisers of the shared task, which uses the baseline features provided by them and a standard logistic regression binary classifier.

As can be seen in Table 1, the results obtained on the development set and the test set are quite similar and coherent, which highlights the robustness of the approach. The results obtained clearly outperform the baseline on the main evaluation metric (F_1^{BAD}). It is worth noting that, on this metric, the



Data set	System	P^{BAD}	R^{BAD}	F_1^{BAD}	P^{GOOD}	R^{GOOD}	F_1^{GOOD}	F_1^w
development set	SBI	31.2%	63.7%	41.9%	88.5%	66.7%	76.1%	69.5%
	SBI+baseline	33.4%	60.9%	43.1%	88.5%	71.1%	78.8%	72.0%
test set	baseline	—	—	16.8%	—	—	88.9%	75.3%
	SBI	30.8%	63.9%	41.5%	88.8%	66.5%	76.1%	69.5%
	SBI+baseline	32.6%	63.6%	43.1%	89.1%	69.5%	78.1%	71.5%

Table 1: Results of the two systems submitted to the WMT 2015 sub-task on word-level MTQE: the one using only sources of bilingual information (SBI) and the one combining these sources of information with the baseline features (SBI+baseline). The table also includes the results of the baseline system proposed by the organisation; in this case only the F_1 scores are provided because, at the time of writing this paper, the rest of metrics remain unpublished.

SBI and SBI+baseline submissions scored first and third among the 16 submissions to the shared task.[8] The submission scoring second obtained very similar results; for F_1^{BAD} it obtained 43.05%, while our submission obtained 43.12%. On the other hand, using the metric F_1^w for comparison, our submissions ranked 10 and 11 in the shared task, although it is worth noting that our system was optimised using only the F_1^{BAD} metric, which is the one chosen by the organisers for ranking submissions.

4 Concluding remarks

In this paper we described the submissions of the *UAlacant* team for the sub-task 2 in the MTQE shared task of the WMT 2015 (word-level MTQE). Our submissions, which were ranked first and third, used online available sources bilingual of information in order to extract relations between the words in the original SL segments and their TL machine translations. The approach employed is aimed at being system-independent, since it only uses resources produced by external systems. In addition, adding new sources of information is straightforward, which leaves considerable room for improvement. In general, the results obtained support the conclusions obtained by Esplà-Gomis et al. (2015b) regarding the feasibility of this approach and its performance.

Acknowledgements

Work partially funded by the Spanish Ministerio de Ciencia e Innovación through project TIN2012-32615 and by the European Commission through project PIAP-GA-2012-324414 (Abu-MaTran). We specially thank Reverso-Softissimo and Prompsit Language Engineering for providing the access to the Reverso Context concordancer, and to the University Research Program for Google Translate that granted us access to the Google Translate service.

References

J.S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. 2011. Algorithms for hyper-parameter optimization. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2546–2554.

J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. 2004. Confidence estimation for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04.

R. O. Duda, P. E. Hart, and D. G. Stork. 2000. *Pattern Classification*. John Wiley and Sons Inc., 2nd edition.

M. Esplà-Gomis, F. Sánchez-Martínez, and M.L. Forcada. 2012. UAlacant: Using online machine translation for cross-lingual textual entailment. In *Proceedings of the 6th International Workshop on Semantic Evaluation*, SemEval 2012), pages 472–476, Montreal, Canada.

M. Esplà-Gomis, F. Sánchez-Martínez, and M. L. Forcada. 2015a. Target-language edit hints in CAT tools based on TM by means of MT. *Journal of Artificial Intelligence Research*, 53:169–222.

M. Esplà-Gomis, F. Sánchez-Martínez, and M. L. Forcada. 2015b. Using on-line available sources of bilingual information for word-level machine translation quality estimation. In *Proceedings of the 18th Annual Conference of the European Associtation for Machine Translation*, pages 19–26, Antalya, Turkey.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. 2009. The WEKA Data Mining Software: an Update. *SIGKDD Explorations*, 11(1):10–18.

L. Specia and R. Soricut. 2013. Quality estimation for machine translation: preface. *Machine Translation*, 27(3-4):167–170.

[8] `http://www.quest.dcs.shef.ac.uk/wmt15_files/results/task2.pdf`

L. Specia, D. Raj, and M. Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50.

R.A. Wagner and M.J. Fischer. 1974. The string-to-string correction problem. *Journal of the ACM*, 21(1):168–173.

Chapter 4

Creation of sources of bilingual information for under-resourced language pairs from multilingual websites

This chapter describes the creation of new SBI for under-resourced language pairs through parallel data crawling from multilingual websites. This research is conducted to confirm our last working hypotheses:

Hypothesis #5: it is possible to create new SBI that enable word-level QE for language pairs with no SBI available using Bitextor to crawl parallel data.

Hypothesis #6: the results obtained in word-level QE for under-resourced language pairs can be improved by using new SBI obtained through parallel data crawling.

As can be seen both hypotheses are similar, but there is an important nuance that differentiates them. Working hypothesis #5 declares the usefulness of the SBI derived from parallel data crawling when there are no other SBI to be used. Whereas, working hypothesis #6 goes a step beyond and predicts that these new SBI should improve the results, even when other SBI are available. Figure 4.1 highlights the research work covered in this chapter and its relations with the rest of the work in this PhD thesis.

The whole research described in this chapter is somehow parallel to the rest of methods developed as part of this PhD thesis, since it provides the basis for applying them to language pairs with few or none SBI available. The focus of the work described in this chapter is on Croatian, which, according to the work by Rehm and

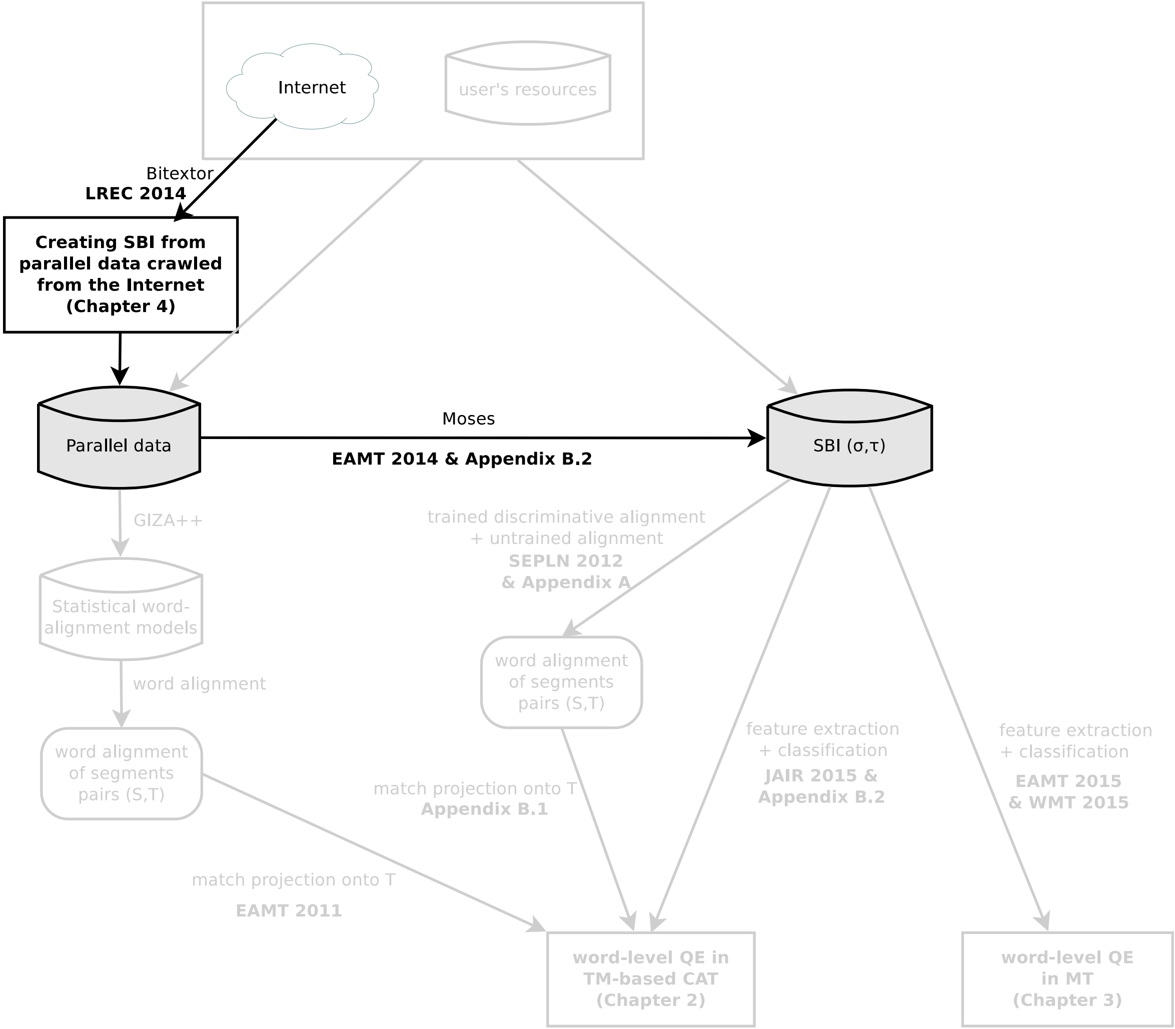

Internet
user's resources
Bitextor
LREC 2014
Creating SBI from parallel data crawled from the Internet (Chapter 4)
Parallel data
Moses
EAMT 2014 & Appendix B.2
SBI (σ,τ)
GIZA++
Statistical word-alignment models
word alignment
word alignment of segments pairs (S,T)
trained discriminative alignment + untrained alignment
SEPLN 2012 & Appendix A
word alignment of segments pairs (S,T)
match projection onto T
Appendix B.1
feature extraction + classification
JAIR 2015 & Appendix B.2
feature extraction + classification
EAMT 2015 & WMT 2015
match projection onto T
EAMT 2011
word-level QE in TM-based CAT (Chapter 2)
word-level QE in MT (Chapter 3)

Figure 4.1: This diagram highlights the research being reported in Chapter 4 (and the publications concerned) on the development of methods that create SBI for under-resourced language pairs by crawling parallel data from multilingual websites, and relates it to the rest of the work reported in this dissertation.

Uszkoreit (2013), figures among the European languages with fewer resources. Two publications are included in this chapter:

- Esplà-Gomis, M., Klubička, F., Ljubešić, N., Ortiz-Rojas, S., Papavassiliou, S., and Prokopidis, P. 2014. Comparing two acquisition systems for automatically building an English–Croatian parallel corpus from multilingual websites. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, p. 1252–1258, Reykjavík, Iceland, May 26–31, 2014. [**Reprinted publication 4.1**]
- Toral, A., Rubino, R., Esplà-Gomis, M., Pirinen, T., Way, A., and Ramírez-Sánchez, G. 2014. Extrinsic evaluation of web-crawlers in machine translation: a case study on Croatian–English for the tourism domain. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, p. 221–224, Dubrovnik, Croatia, June 16–18, 2014. [**Reprinted publication 4.2**]

Reprinted publication 4.1 compares two state-of-the-art multilingual crawlers: Bitextor (Esplà-Gomis and Forcada, 2010), and the ILSP Focused Crawler (Papavassiliou et al., 2013) for the task of crawling an English–Croatian parallel corpus. The experiments conducted in this work confirm the high quality of the parallel corpora crawled in both cases. In addition, it shows a rather small overlap between the data obtained with each of them, which raises the issue that different approaches can obtain corpora of a similar level of quality covering different parts of the websites crawled.

Reprinted publication 4.2 reports an extrinsic evaluation of the same corpus described in reprinted publication 4.1. This extrinsic evaluation consists in training a phrase-based SMT system (Koehn et al., 2003) by using this parallel corpus to train the free/open-source SMT system Moses (Koehn et al., 2007). The MT system is then evaluated for the translation of Croatian into English. MT is one kind of SBI that can be obtained from parallel corpora, although other SBI can be easily obtained (phrase tables, probabilistic bilingual dictionaries, TMs, etc.).

In Appendix B, unpublished experiments are reported on the acquisition and use of SBI aimed at confirming the working hypotheses #5 and #6. These additional experiments could not be included in this chapter because they have not been published in peer-reviewed publications, a requirement imposed by the thesis-by-compilation rules at Universitat d'Alacant. In these unpublished experiments, word-level QE in TM-based CAT is evaluated for two new SBI built on data crawled from the Internet: phrase tables and phrase-based SMT systems. The evaluation is performed for the most under-resourced language pair used in the experiments described in reprinted publication 2.2.1: English–Finnish. The results show that the SBI crawled from the Internet are useful and improve the coverage of our TM-based CAT QE method almost three times when compared to the original results.

Comparing two acquisition systems for automatically building an English–Croatian parallel corpus from multilingual websites

Miquel Esplà-Gomis,[*] Filip Klubička,[†] Nikola Ljubešić,[†]
Sergio Ortiz-Rojas,[‡] Vassilis Papavassiliou,[§] Prokopis Prokopidis[§]

[*]University of Alacant, Alacant (Spain)
mespla@dlsi.ua.es

[†]University of Zagreb, Zagreb (Croatia)
fklubicka@ffzg.hr, nikola.ljubesic@ffzg.hr

[‡]Prompsit Language Engenering, Elx (Spain)
sergio@prompsit.com

[§]Institute for Language and Speech Processing, Athens (Greece)
vpapa@ilsp.gr, prokopis@ilsp.gr

Abstract

In this paper we compare two tools for automatically harvesting bitexts from multilingual websites: *bitextor* and *ILSP-FC*. We used both tools for crawling 21 multilingual websites from the tourism domain to build a domain-specific English–Croatian parallel corpus. Different settings were tried for both tools and 10,662 unique document pairs were obtained. A sample of about 10% of them was manually examined and the success rate was computed on the collection of pairs of documents detected by each setting. We compare the performance of the settings and the amount of different corpora detected by each setting. In addition, we describe the resource obtained, both by the settings and through the human evaluation, which has been released as a high-quality parallel corpus.

Keywords: bitext crawling, parallel corpora, Croatian

1. Introduction

Parallel corpora are a valuable source of cross-lingual knowledge, consisting of collections of text-fragment pairs, usually known as *bitexts* (Harris, 1988), which are mutual translations in different languages. These corpora have been shown to be a useful resource for a wide range of tasks in natural language processing (Melamed, 2001), such as cross-lingual information retrieval (Nie et al., 1999), cross-lingual textual entailment (Mehdad et al., 2011), or word-sense disambiguation (Diab and Resnik, 2002). However, it is in statistical machine translation (SMT) (Koehn, 2010) where the use of parallel corpora is more relevant. The proliferation of parallel-corpora-based methods has raised a growing interest on parallel corpora collection in the last decades.

Many sources of bitexts have been identified: parallel corpora have been built from legal texts, such as the Hansards corpus (Roukos et al., 1995) or the Europarl corpus (Koehn, 2005); translations of software interfaces and documentation, such as KDE4 and OpenOffice (Tiedemann, 2009); or news translated into different languages, such as the SE-Times corpus (Tiedemann, 2009), or the News Commentaries corpus (Bojar et al., 2013), etc.

One of the hugest sources of parallel corpora is the Internet, since there are many websites which are available in two or more languages. Many approaches have been therefore proposed for trying to exploit the Web as a parallel corpus. One of the most complex tasks involved in this problem is parallel document identification. Three main strategies can be found in the literature for parallel document identification in multilingual websites by exploiting:

- similarities in the URLs corresponding to web pages from a web site (Ma and Liberman, 1999; Nie et al., 1999; Resnik and Smith, 2003; Chen et al., 2004; Zhang et al., 2006; Désilets et al., 2008; Esplà-Gomis and Forcada, 2010; San Vicente and Manterola, 2012);
- parallelisms in the structure of HTML files (Nie et al., 1999; Resnik and Smith, 2003; Sin et al., 2005; Shi et al., 2006; Zhang et al., 2006; Désilets et al., 2008; Esplà-Gomis and Forcada, 2010; San Vicente and Manterola, 2012; Papavassiliou et al., 2013); and
- content-similarity techniques (mostly based on bag-of-words overlapping metrics) (Ma and Liberman, 1999; Chen et al., 2004; Zhang et al., 2006; Jiang et al., 2009; Utiyama et al., 2009; Yan et al., 2009; Hong et al., 2010; Sridhar et al., 2011; Antonova and Misyurev, 2011; Barbosa et al., 2012).

In addition to these strategies, other heuristics can be found in the bibliography, such as file size comparison, language markers in the HTML structure, mutual hyper-links between web pages, or images co-occurrence (Papavassiliou et al., 2013). It is usual to combine several of these methods in order to improve the performance.

In this work we use two tools from this bibliography, *ILSP-FC*[1] (Papavassiliou et al., 2013) and *bitextor*[2]

[1] http://nlp.ilsp.gr/redmine/projects/ilsp-fc

[2] http://sf.net/projects/bitextor

(Esplà-Gomis and Forcada, 2010), for harvesting English–Croatian parallel documents from a collection of 21 multilingual websites belonging to the tourism domain. In our experiments, we compare the success rate of these settings to detect parallel documents by manually checking a representative sample of the document pairs obtained by each of them. Additionally, we describe the parallel corpus obtained as a by-product of this evaluation.

1.1. Bitextor

Bitextor is a free/open-source tool for harvesting bitexts from multilingual websites. The newest version of bitextor (version 4.0) is a re-implementation of the tool described by Esplà-Gomis and Forcada (2010). In this version, the techniques based on URL similarity are replaced by new methods based on bag-of-words overlapping. Given a multilingual website and the pair of targeted languages (L_1, L_2) from which the parallel corpus has to be created, bitextor performs the following steps:

1. the website is completely downloaded by means of the tool *HTTrack*,[3] keeping only HTML documents;
2. downloaded documents are preprocessed with *Apache Tika*[4] and *boilerpipe*[5] (Kohlschütter et al., 2010) to normalise the HTML structure and remove boilerplates;
3. duplicate documents (regarding the text, not the structure) are removed, and the language of each file is detected with *LangID* (Lui and Baldwin, 2012),[6] keeping only those documents in L_1 or L_2;
4. bag-of-words overlapping metrics are used to choose a preliminary n-best candidates list for each document;
5. each n-best candidates list is re-ranked by using metrics based on the Levenshtein edit distance between the HTML structure of each pair of documents;
6. the most promising document pairs in the n-best candidates lists are aligned and *hunalign*[7] (Varga et al., 2005) is used to obtain an indicative score regarding the quality of the sentence-alignment between both documents.

1.2. ILSP-FC

ILSP-FC is a modular system that includes components and methods for all the tasks required to acquire domain-specific corpora from the Web. Depending on user-defined configuration, the crawler employs processing workflows for the creation of either monolingual corpora or bilingual collections (i.e. pairs of parallel documents acquired from multilingual web sites). The main modules integrated in ILSP-FC are:

1. page fetcher: adopts a multithreaded crawling implementation in order to ensure concurrent visiting of multiple web pages/hosts.
2. normaliser: parses the structure of each fetched web page and extracts its metadata, detects its encoding and converts it to UTF-8 if required.
3. cleaner: extracts structural information (i.e. title, heading, etc.) and identifies boileplate paragraphs.
4. language identifier: uses the *Cybozu*[8] library to detect the main language of a document, as well as paragraphs in a language different from the main one.
5. link extractor: examines the anchor text of the extracted links and ranks them by the probability that a link from a page points to a candidate translation of this page, with the purpose of forcing the crawler to visit candidate translations first.
6. de-duplicator: checks each document against all others and identifies (near-)duplicates by comparing the quantized word frequencies and the paragraphs of each pair of candidate duplicate documents;
7. pair detector: examines each document against all others and identifies pairs of documents that could be considered parallel. Its main methods are based on URL similarity, co-occurrences of images with the same filename in two documents, and the documents' structural similarity.

2. Experimental settings

Our English–Croatian corpus is built from the collection of 21 multilingual websites listed in Table 1. These websites were handpicked from a list of 100 most bitext-productive multilingual websites from the Croatian top-level domain. The list of the most productive multilingual websites was obtained by calculating the website frequency distribution in the *hrenWaC* corpus[9] (Tiedemann, 2009), a side-product of the *hrWaC* Croatian web corpus (Ljubešić and Erjavec, 2011). Our future plans cover combining the procedure of top-level domain crawling for bitext-hotspot identification and bilingual focused crawling of the bitext hotspots for obtaining parallel data.

In our experiments, two different configurations were tried for ILSP-FC:

- *all*: It includes all the pairs detected by the tool (i.e. default configuration);
- *reliable*: It includes a subset of the *all* configuration where only those pairs identified through image co-occurrences and high-structural similarity are kept;

and four were tried for bitextor:

- *10-best*: 10-best candidate lists are used to get the pairs of documents;
- *1-best*: 1-best candidate lists are used to get the pairs of documents; this setting is more strict than *10-best*, since it only aligns documents which are mutual best candidates;

[3] http://www.httrack.com/

[4] http://tika.apache.org/

[5] http://code.google.com/p/boilerpipe/

[6] https://github.com/saffsd/langid.py

[7] http://mokk.bme.hu/resources/hunalign/

[8] http://code.google.com/p/language-detection/

[9] http://nlp.ffzg.hr/resources/corpora/hrenwac/

URL	description
`http://www.adria-bol.hr/`	Website of a tourist agency based in the city of Bol
`http://www.animafest.hr/`	Portal of the World Festival of Animated Film in Zagreb
`http://bol.hr/`	Tourism portal of the city of Bol
`http://www.burin-korcula.hr/`	Website of Burin, a private tourist agency Korula island
`http://www.camping.hr/`	Website of the Croatian Camping Union (CCU)
`http://www.dalmatia.hr/`	Official tourism portal of Dalmatia Country
`http://dubrovnik-festival.hr/`	Website of the Dubrovnik Summer Festival
`http://www.events.hr/`	Croatian online travel agent
`http://www.galileo.hr/`	Croatian online travel agent
`http://hhi.hr/`	Hydrographic Institute of the Republic of Croatia
`http://www.istra.hr/`	Official tourism portal of Istria
`http://www.kvarner.hr/`	Official tourism portal of Kvarner County
`http://plavalaguna.hr`	Website of the hotel company *Laguna Porec*
`http://www.liburnia.hr/`	Website of the hotel company *Liburnia Riviera Hotels*
`http://m.pulainfo.hr/`	Tourism portal of the city of Pula
`http://www.portauthority.hr/`	Website of the Croatian Association of Port Autorities
`http://www.putomania.com.hr`	Portal about travelling around the world
`http://www.tzg-rab.hr/`	Tourism portal about Rab island
`http://tzgrovinj.hr/`	Official tourism portal of Rovinj-Rovigno
`http://www.uniline.hr/`	Festival of urban culture
`http://urbanfestival.blok.hr/`	On-line reservation of accommodation in Croatia

Table 1: List of processed websites including the URL and a short description

- *10-best-filtered*: The same than *10-best*, but those pairs of documents with a segment-alignment score (provided by hunalign) under 0.3 are discarded;
- *1-best-filtered*: The same than *1-best*, but those pairs of documents with a segment-alignment score under 0.3 are discarded.

For these settings, we computed the success ratio obtained for identifying parallel documents by manually verifying a sample of the document pairs obtained. In addition to this quality evaluation, we wanted to obtain a quantitative measure of the amount of data crawled by each setting. However, using only the amount of parallel documents detected to this end presents a problem: bitextor and ILSP-FC adopt different strategies for discarding duplicates. While ILSP-FC discards (near-)duplicate documents, bitextor only discards documents containing exactly the same text. As a result, bitextor retrieves much more document pairs than ILSP-FC, but the degree of redundancy is much higher. In order to perform a fair comparison between both tools, we decided to measure the number of unique aligned segments and, therefore, to reduce the impact of redundancy in the data obtained by bitextor. To perform the alignment of the document pairs at the segment level, both corpora were further segmented into sentences[10] and tokenised using the scripts[11] and included in the Moses statistical machine translation toolkit (Koehn et al., 2007). Then, the tool hunalign was used for aligning the segments. Finally, segment pairs with a score lower than 0 were discarded.

3. Results and discussion

The pairs of documents detected by each setting were merged in a pool containing 10,662 unique pairs of documents. As expected, we observed a high degree of overlapping between the settings of the same tool.[12] However, only 8.5% of the document pairs in *all* were also in *10best*. This divergence is due to the different methods used by each tool to crawl the websites and to detect parallel documents, and suggests that they could be combined to obtain a bigger corpus. Table 2 shows the total amount of document pairs obtained with each setting, as well as the number of unique segments contained in these documents both in English and Croatian. The last column of the table contains the number of unique segment pairs obtained after aligning the collection of document pairs obtained with the tool hunalign.[13] It is worth noting that the relative difference between the numbers of parallel documents obtained by each setting is much higher than the relative difference between the numbers of unique aligned segment pairs. This confirms the idea that the number of document pairs is not an appropriate metric to check the amount of data obtained with each tool, as mentioned in Section 2.

From the pool of document pairs, a sample of 1,129 (about 10%) was randomly picked and checked, obtaining a total

[10] Both Bitextor and FC split the text in a document by using the HTML tags in it. However, it is possible to have pieces of text longer than a segment, so a second segmentation process is required.

[11] `https://github.com/moses-smt/mosesdecoder/blob/master/scripts/ems/support/split-sentences.perl` and `https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl`

[12] As already mentioned, settings *10best-filtered*, *1best-filtered*, and *reliable* are sub-sets of *10best*, *1best*, and *all*, respectively; in addition, 97.9% of the pairs of documents in *1best* also appeared in *10best*.

[13] All the data provided in Table 2 regarding segments was lowercased before removing duplicates in order to minimise the redundancy.

tool	setting	aligned documents	unique segments		unique aligned segment pairs
			English	Croatian	
focused crawler	all	3,294	46,226	47,370	40,431
	reliable	2,406	37,986	38,772	32,544
bitextor	10best	7,787	54,859	46,794	50,338
	10best-filtered	5,056	49,406	43,972	46,242
	1best	4,232	41,318	40,703	37,727
	1best-filtered	3,758	40,078	39,542	36,834

Table 2: Amount of document pairs obtained with each of the two settings of ILSP-FC, and for the four settings of bitextor. The table also reports the number of unique lowercased segments from the aligned documents both in English and in Croatian, and the number of unique lowercased aligned segment pairs obtained after aligning all these documents.

of 831 pairs confirmed as parallel documents by the human evaluators. Table 3 shows the success rates obtained by each setting when identifying parallel documents. These results confirm that, as expected, the *reliable* setting provides better precision than *all* for ILSP-FC, while the settings *1best* and *1best-filtered* are the most successful for bitextor. In a general comparison, *1best-filtered* overcomes all the other settings in terms of success rate. Another interesting detail is that the fraction of parallel documents in the whole sample is 73.6%, which is lower than the success rate obtained by each setting. This is due to the fact that the intersection of the pairs of documents obtained by all settings contains more parallel documents than non-parallel documents. In order to examine the intersection of each setting against the others and check the contribution of each setting to the resulting corpus, a similarity measurement was performed between the sub-corpora obtained with each setting. Thus, Table 4 shows the Jaccard index (Chakrabarti, 2003, Chapter 3) between the collections of aligned segment pairs obtained with each setting. Additionally, the last column of this table reports the Jaccard index between the corpus obtained with each setting and the resulting corpus, this is, the part of this corpus covered by each setting. These results show that the pair detectors integrated in these two tools could be considered complementary. For instance, the accuracy rates of the *reliable* setting of ILSP-FC and the *1best-filtered* of bitextor are 90.76% and 94.79% respectively while only 13,44% of the delivered unique segment pairs are common. Hence, it seems logical to use both tools in parallel to maximise the amount of parallel data collected from a collection of websites. Comparing the results regarding the Jaccard index of each setting with the whole corpus obtained, we can conclude that the contribution of both ILSP-FC and bitextor is quite balanced.

4. Error analysis

We devoted some time to check which were the main errors made by each tool when detecting parallel documents and some patterns were observed. Typical errors were:

- *content similarity*: Some of the websites crawled were prone to contain very similar web pages. For example, in the case of hotel chains, it is usual to find web pages about different hotels, where most of the text is the same and only a few data (name, address, number of rooms, etc.) changes. These similarities in the content caused many wrong document alignments, which were more usual in the case of bitextor, which does not remove near-duplicate documents. It is worth noting that these errors at the level of document alignment are not so severe when the corpus is aligned at segment level, since most of the aligned segment pairs are correct.
- *URL similarity*: In the case of ILSP-FC, websites keeping a highly similar URL structure caused also wrong alignments, since one of the strategies adopted by this tool is to compare URLs ignoring the differences in the content of the pages.

tool	setting	success rate
focused crawler	all	73.86%
	reliable	90.76%
bitextor	10best	74.70%
	10best-filtered	83.57%
	1best	92.68%
	1best-filtered	94.79%

Table 3: Results on the manual revision of detected parallel documents. For each setting, number of pairs of documents detected which were confirmed o be parallel.

5. Resulting corpus

Two parallel English–Croatian corpora were obtained as a result of this work: a general corpus resulting from the union of all the 10,662 pairs of documents obtained by each setting, and a human-verified corpus resulting of the compilation of all the 831 documents confirmed as parallel by the human evaluators. These corpora are available at `http://redmine.abumatran.eu/projects/en-hr-tourism-corpus` aligned at the segment level[14] and formatted following the TMX stan-

[14] The alignment was performed following the methodology described in Section 2.

		Jaccard index between aligned corpora						
		focused		*bitextor*				*merged*
		all	**reliable**	**10best**	**10best-filtered**	**1best**	**1best-filtered**	
focused crawler	all	—	70.84%	10.93%	11.38%	12.04%	12.06%	46.46%
	reliable	—	—	11.69%	12.28%	13.19%	13.22%	37.40%
bitextor	10best	—	—	—	86.62%	68.28%	67.12%	57.84%
	10best-filtered	—	—	—	—	72.23%	73.40%	53.14%
	1best	—	—	—	—	—	95.34%	43.35%
	1best-filtered	—	—	—	—	—	—	42.33%

Table 4: Jaccard index measuring the similarity between the different collections of unique segment pairs obtained with each setting. The final column measures the Jaccard index of each setting with the *merged* corpus obtained when producing the union of all the settings.

dard.[15] In addition, a field *prop*[16] was added to each unit in the TMX file containing a comma-separated list with the names of the settings which produced it. This information is aimed at allowing to extract customised sub-corpora with different degrees of quality, depending on the settings included. After alignment, we obtained 87,024 aligned segments for the general corpus, and 9,387 for the human-verified corpus.

6. Concluding remarks

In this work we compared two tools for automatically crawling parallel corpora from multilingual websites: Focused Crawler and Bitextor. We used both tools for crawling 21 websites in the tourism domain in order to build an English–Croatian domain-specific corpus. We used several settings for crawling with each tool in order to compare them in terms of amount of parallel data obtained and precision in parallel document crawling. Our experiments proved that both tools can obtain similar precision and amount of data depending on the setting chosen. In addition, we proved that both tools obtain parallel data from different parts of the websites and, therefore, combining the corpora obtained by them allows us to mine parallel documents more exhaustively.

We finally obtained a parallel corpus consisting of 10,662 pairs of documents, which, after segment alignment, resulted in a collection of 87,024 unique pairs of segments. In addition, the human verification performed for evaluating precision allowed us to produce a smaller high-quality parallel corpus consisting of 831 pairs of documents, which were manually verified as parallel documents. After aligning this second corpus at the level of segments, we obtained 9,387 unique pairs of segments.

7. Acknowledgements

The research leading to these results has received funding from the European Commission through project PIAP-GA-2012-324414 (Abu-MaTran) and the Spanish government through project TIN2012-32615.

8. References

Antonova, Alexandra and Misyurev, Alexey. (2011). Building a web-based parallel corpus and filtering out machine-translated text. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 136–144, Portland, Oregon. Association for Computational Linguistics.

Barbosa, Luciano, Rangarajan Sridhar, Vivek Kumar, Yarmohammadi, Mahsa, and Bangalore, Srinivas. (2012). Harvesting parallel text in multiple languages with limited supervision. In *Proceedings of COLING 2012*, pages 201–214, Mumbai, India.

Bojar, Ondřej, Buck, Christian, Callison-Burch, Chris, Federmann, Christian, Haddow, Barry, Koehn, Philipp, Monz, Christof, Post, Matt, Soricut, Radu, and Specia, Lucia. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the 8th Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria.

Chakrabarti, Soumen. (2003). *Mining the Web: Discovering knowledge from hypertext data*. Morgan Kaufmann.

Chen, Jisong, Chau, Rowena, and Yeh, Chung-Hsing. (2004). Discovering parallel text from the world wide web. In *Proceedings of the Second Workshop on Australasian Information Security, Data Mining and Web Intelligence, and Software Internationalisation*, volume 32 of *ACSW Frontiers'04*, pages 157–161, Dunedin, New Zealand.

Désilets, Alain, Farley, Benoit, Stojanovic, M, and Patenaude, G. (2008). WeBiText: Building large heterogeneous translation memories from parallel web content. In *Proceedings of Translating and the Computer*, pages 27–28, London, UK.

Diab, Mona and Resnik, Philip. (2002). An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL'02, pages 255–262, Philadelphia, Pennsylvania.

Esplà-Gomis, Miquel and Forcada, Mikel L. (2010). Combining content-based and URL-based heuristics to harvest aligned bitexts from multilingual sites with bitextor. *The Prague Bulletin of Mathematical Linguistics*, 93:77–

[15] http://www.gala-global.org/oscarStandards/tmx/tmx14b.html

[16] http://www.gala-global.org/oscarStandards/tmx/tmx14b.html#prop

86.

Harris, Brian. (1988). Bi-text, a new concept in translation theory. *Language Monthly*, 54:8–10.

Hong, Gumwon, Li, Chi-Ho, Zhou, Ming, and Rim, Hae-Chang. (2010). An empirical study on web mining of parallel data. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING'10, pages 474–482, Beijing, China. Association for Computational Linguistics.

Jiang, Long, Yang, Shiquan, Zhou, Ming, Liu, Xiaohua, and Zhu, Qingsheng. (2009). Mining bilingual data from the web with adaptively learnt patterns. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, volume 2 of *ACL'09*, pages 870–878, Suntec, Singapore.

Koehn, Philipp, Hoang, Hieu, Birch, Alexandra, Callison-Burch, Chris, Federico, Marcello, Bertoldi, Nicola, Cowan, Brooke, Shen, Wade, Moran, Christine, Zens, Richard, Dyer, Chris, Bojar, Ondřej, Constantin, Alexandra, and Herbst, Evan. (2007). Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL'07, pages 177–180, Prague, Czech Republic.

Koehn, Philipp. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the X Machine Translation Summit*, pages 79–86, Phuket, Thailand.

Koehn, Philipp. (2010). *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

Kohlschütter, Christian, Fankhauser, Peter, and Nejdl, Wolfgang. (2010). Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 441–450, New York, NY, USA.

Ljubešić, Nikola and Erjavec, Tomaž. (2011). hrWaC and slWac: Compiling web corpora for Croatian and Slovene. In Habernal, Ivan and Matousek, Václav, editors, *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, Lecture Notes in Computer Science, pages 395–402. Springer.

Lui, Marco and Baldwin, Timothy. (2012). Langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, ACL'12, pages 25–30, Jeju Island, Korea.

Ma, Xiaoyi and Liberman, Mark. (1999). Bits: A method for bilingual text search over the web. In *Machine Translation Summit VII*, pages 538–542, Singapore, Singapore.

Mehdad, Yashar, Negri, Matteo, and Federico, Marcello. (2011). Using bilingual parallel corpora for cross-lingual textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1 of *HLT'11*, pages 1336–1345, Portland, Oregon. Association for Computational Linguistics.

Melamed, Dan I. (2001). *Empirical methods for exploiting parallel texts*. MIT Press.

Nie, Jian-Yun, Simard, Michel, Isabelle, Pierre, and Durand, Richard. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'99, pages 74–81, Berkeley, California, USA. ACM.

Papavassiliou, Vassilis, Prokopidis, Prokopis, and Thurmair, Gregor. (2013). A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 43–51, Sofia, Bulgaria. Association for Computational Linguistics.

Resnik, Philip and Smith, Noah A. (2003). The Web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Roukos, Salim, Graff, David, and Melamed, Dan. (1995). Hansard French/English. Linguistic Data Consortium. Philadelphia, USA.

San Vicente, Iaki and Manterola, Iker. (2012). PaCo2: A fully automated tool for gathering parallel corpora from the web. In Chair), Nicoletta Calzolari (Conference, Choukri, Khalid, Declerck, Thierry, Doan, Mehmet Uur, Maegaard, Bente, Mariani, Joseph, Odijk, Jan, and Piperidis, Stelios, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, LREC'1, Istanbul, Turkey. European Language Resources Association (ELRA).

Shi, Lei, Niu, Cheng, Zhou, Ming, and Gao, Jianfeng. (2006). A DOM tree alignment model for mining parallel data from the web. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL'06, pages 489–496, Sydney, Australia.

Sin, Chunyu, Liu, Xiaoyue, Sin, KingKui, and Webster, Jonathan J. (2005). Harvesting the bitexts of the laws of Hong Kong from the web. In *Proceedings of the Fifth Workshop on Asian Language Resources and First Symposium on Asian Language Resources Network*, pages 71–78, Jeju, South Corea.

Sridhar, Vivek Kumar Rangarajan, Barbosa, Luciano, and Bangalore, Srinivas. (2011). A scalable approach to building a parallel corpus from the web. In *Interspeech*, pages 2113–2116, Florence, Italy.

Tiedemann, Jörg. (2009). News from OPUS - a collection of multilingual parallel corpora with tools and interfaces. In Nicolov, N., Bontcheva, K., Angelova, G., and Mitkov, R., editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Utiyama, Masao, Kawahara, Daisuke, Yasuda, Keiji, and Sumita, Eiichiro. (2009). Mining parallel texts from mixed-language web pages. In *Proceedings of the XII Machine Translation Summit*, Ottawa, Ontario, Canada.

Varga, Dániel, Németh, László, Halácsy, Péter, Kornai, András, Trón, Viktor, and Nagy, Viktor. (2005). Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing*, pages 590–596, Borovets, Bulgaria.

Yan, Zhenxiang, Feng, Yanhui, Hong, Yu, and Yao, Jianmin. (2009). Parallel sentences mining from the web. *Journal of Computational Information Systems*, 6:1633–1641.

Zhang, Ying, Wu, Ke, Gao, Jianfeng, and Vines, Phil. (2006). Automatic acquisition of Chinese–English parallel corpus from the web. In Lalmas, Mounia, MacFarlane, Andy, Rger, Stefan, Tombros, Anastasios, Tsikrika, Theodora, and Yavlinsky, Alexei, editors, *Advances in Information Retrieval*, volume 3936 of *Lecture Notes in Computer Science*, pages 420–431. Springer Berlin Heidelberg.

Extrinsic Evaluation of Web-Crawlers in Machine Translation: a Case Study on Croatian–English for the Tourism Domain*

Antonio Toral[†], **Raphael Rubino**[⋆], **Miquel Esplà-Gomis**[‡], **Tommi Pirinen**[†], **Andy Way**[†], **Gema Ramírez-Sánchez**[⋆]
[†] NCLT, School of Computing, Dublin City University, Ireland
{atoral,tpirinen,away}@computing.dcu.ie
[⋆] Prompsit Language Engineering, S.L., Elche, Spain
{rrubino,gramirez}@prompsit.com
[‡] Dep. Llenguatges i Sistemes Informàtics, Universitat d'Alacant, Spain
mespla@dlsi.ua.es

Abstract

We present an extrinsic evaluation of crawlers of parallel corpora from multilingual web sites in machine translation (MT). Our case study is on Croatian to English translation in the tourism domain. Given two crawlers, we build phrase-based statistical MT systems on the datasets produced by each crawler using different settings. We also combine the best datasets produced by each crawler (union and intersection) to build additional MT systems. Finally we combine the best of the previous systems (union) with general-domain data. This last system outperforms all the previous systems built on crawled data as well as two baselines (a system built on general-domain data and a well known online MT system).

1 Introduction

Along with the addition of new member states to the European Union (EU), the commitment with multilingualism in the EU is strengthened to give support to new languages. This is the case of Croatia, the last member to join the EU in July 2013, and of the Croatian language, which became then an official language of the EU.

Croatian is the third official South Slavic language in the EU along with Bulgarian and Slovene. Other surrounding languages (e.g. Serbian and Bosnian), although still not official in the EU, belong also to the same language family and are the official languages of candidate member states, thus being also of strategic interest for the EU.

We focus on providing machine translation (MT) support for Croatian and other South Slavic languages using and producing publicly available resources. Following our objectives, we developed a general-domain MT system for Croatian–English and made it available online on the day Croatia joined the EU. It is, to the best of our knowledge, the first available MT system for this language pair based on free/open-source technologies.

New languages in the EU like Croatian can benefit from MT to speed up the flow of information from and into other EU languages. While this is the case for most types of content it is especially true for official documentation and for content in particular strategic sectors.

Tourism is one of the most important economic sectors in Croatia. It represented 15.4% of Croatia's gross domestic product in 2012 (up from 14.4% in 2011).[1] With almost 12 million foreign tourists visiting Croatia annually, the tourism sector results in income of 6.8 billion euro.

The increasing number of tourists in Croatia makes tourism a relevant domain for MT in order to provide them with quick and up-to-date information about the country they are visiting. Although most visitors come from non-English speaking countries,[2] English is frequently used as a lingua franca. This observation led us to our first approach to support the Croatian tourism sec-

*The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran).

[1] http://www.eubusiness.com/news-eu/croatia-economy.nrl

[2] According to the site croatia.eu, top emitting countries are Germany (24.2%), Slovenia (10.8%), Austria (8.9%), Italy (7.9%), Czech Republic (7.9%), etc.

tor: to provide MT adapted to the tourism domain from Croatian into English. Later, we will provide MT in the visitors' native languages, i.e. German, Slovene, etc.

We take advantage of a recent work that crawled parallel data for Croatian–English in the tourism domain (Esplà-Gomis et al., 2014). Several datasets were acquired by using two systems for crawling parallel data with a number of settings. In this paper we assess these datasets by building MT systems on them and checking the resulting translation performance. Hence, this work can be considered as an extrinsic evaluation of these crawlers (and their settings) in MT.

Besides building MT systems upon the domain-specific crawled data, we study the concurrent exploitation of domain-specific and general-domain data, with the aim of improving the overall performance and coverage of the system. From this perspective, our case study falls in the area of domain adaptation of MT, following previous works in domains such as labour legislation and natural environment for English–French and English–Greek (Pecina et al., 2012) and automotive for German to Italian and French (Läubli et al., 2013).

The rest of the paper is organised as follows. Section 2 presents the crawled datasets used in this study and details the processing undertaken to prepare them for MT. Section 3 details the different MT systems built. Section 4 shows and comments the results obtained. Finally, Section 5 draws conclusions and outlines future lines of work.

2 Crawled Datasets

Datasets were crawled using two crawlers: ILSP Focused Crawler (FC) (Papavassiliou et al., 2013) and Bitextor (Esplà-Gomis et al., 2010). The detection of parallel documents was carried out with two settings for each crawler: 10best and 1best for Bitextor and reliable and all for FC (see (Esplà-Gomis et al., 2014) for further details). It is worth mentioning that reliable and 1best are subsets of all and 10best, respectively. These subsets were obtained with a more strict configuration of each crawler and, therefore, are expected to contain higher quality parallel text. In addition, a set of parallel segments was obtained by aligning only those pairs of documents which were checked manually by two native speakers of Croatian.

Both Bitextor and FC segment the documents aligned by using the HTML tags. These segments were re-segmented in shorter segments and tokenised with the sentence splitter and tokeniser included in the Moses toolkit.[3]

The resulting segments were then aligned with Hunalign (Varga et al., 2005), using the option `realign`, which provides a higher quality alignment by aligning the output of the first alignment. The documents from each website were concatenated prior to aligning them using tags (`<p>`) to mark document boundaries. Aligning multiple documents at once allows Hunalign to build a larger dictionary for alignment while ensuring that only segments belonging to the same document pair are aligned to each other. The resulting pairs of segments were filtered to remove those with a confidence score lower than 0.4.[4]

From the aligned segments coming from manually checked document pairs we remove duplicate segments. We only keep pairs of segments with confidence score higher than 1.[5] These segments are randomised and we keep two sets, one of 825 segmens for the development set and one of 816 segments for the test set.

From the other 4 datasets, those obtained with the different settings of the two crawlers (1best, 10best, all and reliable), duplicate pairs of segments were also removed. Pairs of segments appearing either in the test or development set were also removed. The remaining pairs of segments are kept and will be used for training MT systems.

Apart from the domain-specific crawled data we use additional general-domain (gen) data gathered from several sources of Croatian–English parallel data: hrenWaC,[6] SETimes[7] and TED Talks.[8] These three datasets are concatenated and will be used to build a baseline MT system.

Table 1 presents statistics (number of sentence pairs, number of tokens and number of unique tokens in source (Croatian) and target (English) language) of the previously introduced parallel datasets for Croatian–English. The table shows

[3] `https://github.com/moses-smt/mosesdecoder`

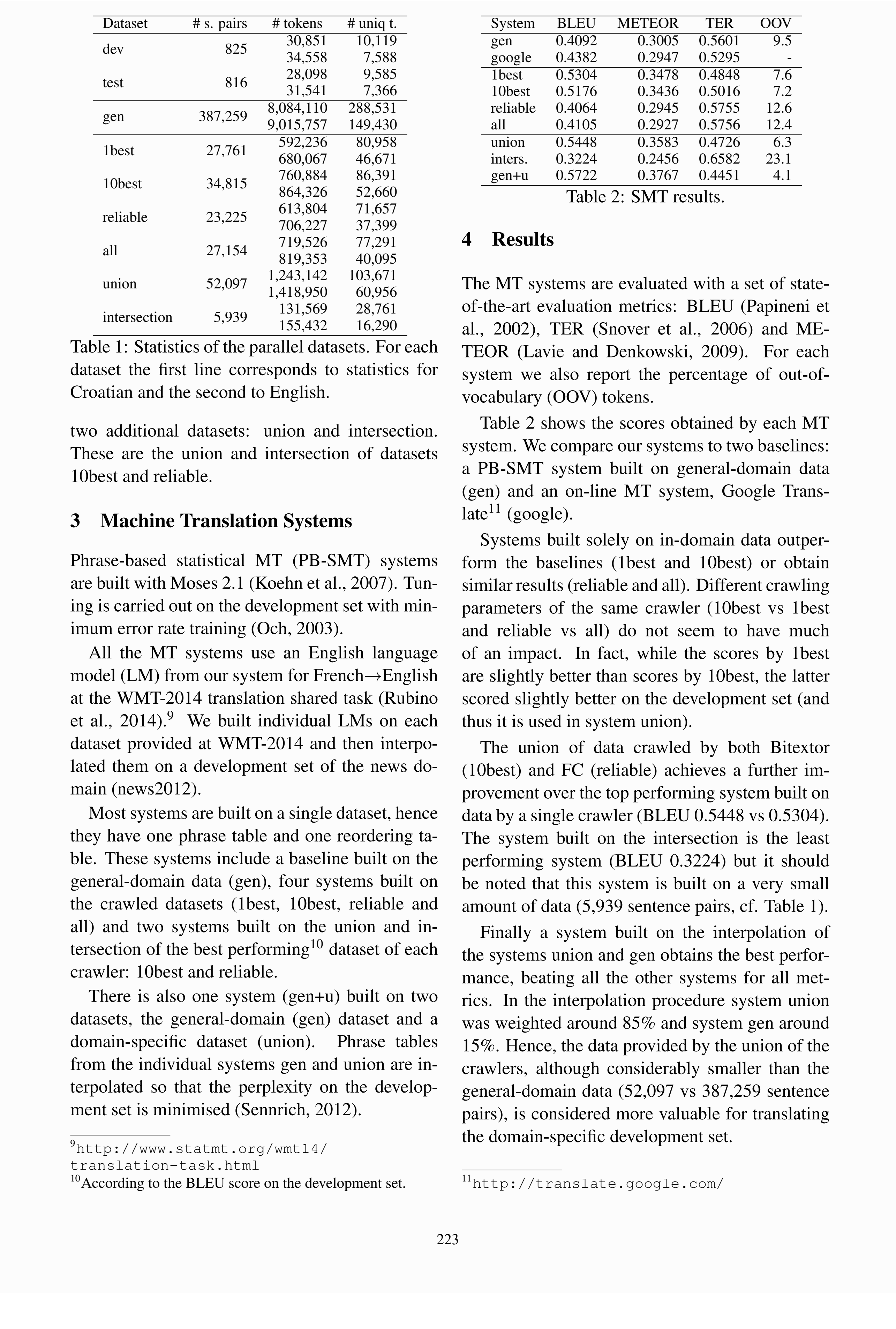
[4] Manual evaluation for English, French and Greek concluded that 0.4 was an adequate threshold for Hunalign's confidence score (Pecina et al., 2012).

[5] While segment pairs with score above 0.4, as shown above, are deemed to be of reasonable quality for training, we raise the threshold to 1 for test and development data.

[6] `http://nlp.ffzg.hr/resources/corpora/hrenwac/`

[7] `http://nlp.ffzg.hr/resources/corpora/setimes/`

[8] `http://zeljko.agic.me/resources/`



Dataset	# s. pairs	# tokens	# uniq t.
dev	825	30,851 34,558	10,119 7,588
test	816	28,098 31,541	9,585 7,366
gen	387,259	8,084,110 9,015,757	288,531 149,430
1best	27,761	592,236 680,067	80,958 46,671
10best	34,815	760,884 864,326	86,391 52,660
reliable	23,225	613,804 706,227	71,657 37,399
all	27,154	719,526 819,353	77,291 40,095
union	52,097	1,243,142 1,418,950	103,671 60,956
intersection	5,939	131,569 155,432	28,761 16,290

Table 1: Statistics of the parallel datasets. For each dataset the first line corresponds to statistics for Croatian and the second to English.

two additional datasets: union and intersection. These are the union and intersection of datasets 10best and reliable.

3 Machine Translation Systems

Phrase-based statistical MT (PB-SMT) systems are built with Moses 2.1 (Koehn et al., 2007). Tuning is carried out on the development set with minimum error rate training (Och, 2003).

All the MT systems use an English language model (LM) from our system for French→English at the WMT-2014 translation shared task (Rubino et al., 2014).[9] We built individual LMs on each dataset provided at WMT-2014 and then interpolated them on a development set of the news domain (news2012).

Most systems are built on a single dataset, hence they have one phrase table and one reordering table. These systems include a baseline built on the general-domain data (gen), four systems built on the crawled datasets (1best, 10best, reliable and all) and two systems built on the union and intersection of the best performing[10] dataset of each crawler: 10best and reliable.

There is also one system (gen+u) built on two datasets, the general-domain (gen) dataset and a domain-specific dataset (union). Phrase tables from the individual systems gen and union are interpolated so that the perplexity on the development set is minimised (Sennrich, 2012).

System	BLEU	METEOR	TER	OOV
gen	0.4092	0.3005	0.5601	9.5
google	0.4382	0.2947	0.5295	-
1best	0.5304	0.3478	0.4848	7.6
10best	0.5176	0.3436	0.5016	7.2
reliable	0.4064	0.2945	0.5755	12.6
all	0.4105	0.2927	0.5756	12.4
union	0.5448	0.3583	0.4726	6.3
inters.	0.3224	0.2456	0.6582	23.1
gen+u	0.5722	0.3767	0.4451	4.1

Table 2: SMT results.

4 Results

The MT systems are evaluated with a set of state-of-the-art evaluation metrics: BLEU (Papineni et al., 2002), TER (Snover et al., 2006) and METEOR (Lavie and Denkowski, 2009). For each system we also report the percentage of out-of-vocabulary (OOV) tokens.

Table 2 shows the scores obtained by each MT system. We compare our systems to two baselines: a PB-SMT system built on general-domain data (gen) and an on-line MT system, Google Translate[11] (google).

Systems built solely on in-domain data outperform the baselines (1best and 10best) or obtain similar results (reliable and all). Different crawling parameters of the same crawler (10best vs 1best and reliable vs all) do not seem to have much of an impact. In fact, while the scores by 1best are slightly better than scores by 10best, the latter scored slightly better on the development set (and thus it is used in system union).

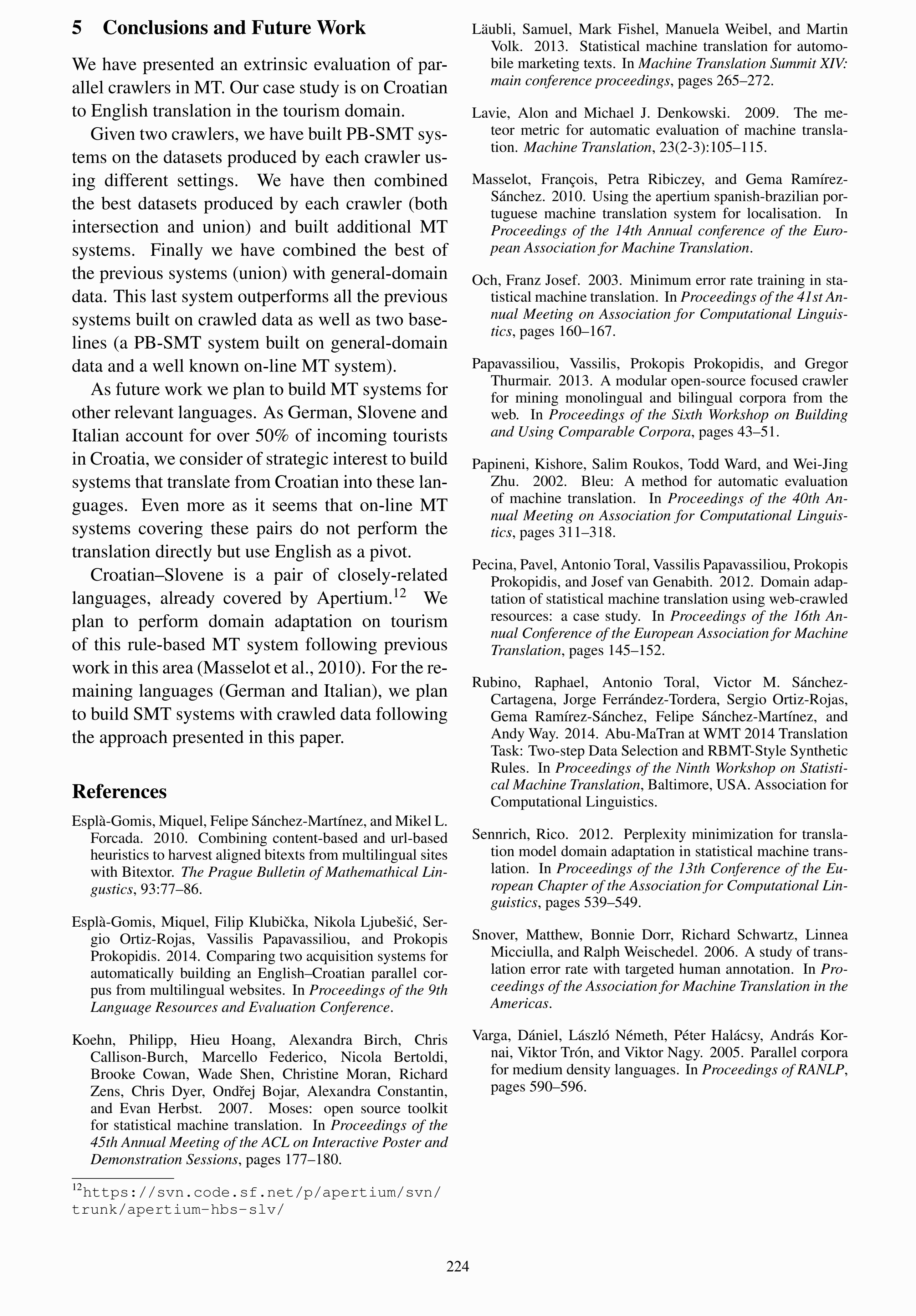
The union of data crawled by both Bitextor (10best) and FC (reliable) achieves a further improvement over the top performing system built on data by a single crawler (BLEU 0.5448 vs 0.5304). The system built on the intersection is the least performing system (BLEU 0.3224) but it should be noted that this system is built on a very small amount of data (5,939 sentence pairs, cf. Table 1).

Finally a system built on the interpolation of the systems union and gen obtains the best performance, beating all the other systems for all metrics. In the interpolation procedure system union was weighted around 85% and system gen around 15%. Hence, the data provided by the union of the crawlers, although considerably smaller than the general-domain data (52,097 vs 387,259 sentence pairs), is considered more valuable for translating the domain-specific development set.

[9] http://www.statmt.org/wmt14/translation-task.html

[10] According to the BLEU score on the development set.

[11] http://translate.google.com/



5 Conclusions and Future Work

We have presented an extrinsic evaluation of parallel crawlers in MT. Our case study is on Croatian to English translation in the tourism domain.

Given two crawlers, we have built PB-SMT systems on the datasets produced by each crawler using different settings. We have then combined the best datasets produced by each crawler (both intersection and union) and built additional MT systems. Finally we have combined the best of the previous systems (union) with general-domain data. This last system outperforms all the previous systems built on crawled data as well as two baselines (a PB-SMT system built on general-domain data and a well known on-line MT system).

As future work we plan to build MT systems for other relevant languages. As German, Slovene and Italian account for over 50% of incoming tourists in Croatia, we consider of strategic interest to build systems that translate from Croatian into these languages. Even more as it seems that on-line MT systems covering these pairs do not perform the translation directly but use English as a pivot.

Croatian–Slovene is a pair of closely-related languages, already covered by Apertium.[12] We plan to perform domain adaptation on tourism of this rule-based MT system following previous work in this area (Masselot et al., 2010). For the remaining languages (German and Italian), we plan to build SMT systems with crawled data following the approach presented in this paper.

References

Esplà-Gomis, Miquel, Felipe Sánchez-Martínez, and Mikel L. Forcada. 2010. Combining content-based and url-based heuristics to harvest aligned bitexts from multilingual sites with Bitextor. *The Prague Bulletin of Mathemathical Linguistics*, 93:77–86.

Esplà-Gomis, Miquel, Filip Klubička, Nikola Ljubešić, Sergio Ortiz-Rojas, Vassilis Papavassiliou, and Prokopis Prokopidis. 2014. Comparing two acquisition systems for automatically building an English–Croatian parallel corpus from multilingual websites. In *Proceedings of the 9th Language Resources and Evaluation Conference*.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180.

Läubli, Samuel, Mark Fishel, Manuela Weibel, and Martin Volk. 2013. Statistical machine translation for automobile marketing texts. In *Machine Translation Summit XIV: main conference proceedings*, pages 265–272.

Lavie, Alon and Michael J. Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2-3):105–115.

Masselot, François, Petra Ribiczey, and Gema Ramírez-Sánchez. 2010. Using the apertium spanish-brazilian portuguese machine translation system for localisation. In *Proceedings of the 14th Annual conference of the European Association for Machine Translation*.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167.

Papavassiliou, Vassilis, Prokopis Prokopidis, and Gregor Thurmair. 2013. A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 43–51.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

Pecina, Pavel, Antonio Toral, Vassilis Papavassiliou, Prokopis Prokopidis, and Josef van Genabith. 2012. Domain adaptation of statistical machine translation using web-crawled resources: a case study. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 145–152.

Rubino, Raphael, Antonio Toral, Victor M. Sánchez-Cartagena, Jorge Ferrández-Tordera, Sergio Ortiz-Rojas, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, and Andy Way. 2014. Abu-MaTran at WMT 2014 Translation Task: Two-step Data Selection and RBMT-Style Synthetic Rules. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA. Association for Computational Linguistics.

Sennrich, Rico. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 539–549.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and Ralph Weischedel. 2006. A study of translation error rate with targeted human annotation. In *Proceedings of the Association for Machine Translation in the Americas*.

Varga, Dániel, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of RANLP*, pages 590–596.

[12] `https://svn.code.sf.net/p/apertium/svn/trunk/apertium-hbs-slv/`

Chapter 5

Concluding remarks

This chapter summarises the main contributions of this PhD thesis to the state of the art in translation QE and outlines future research lines that may be followed in order to extend the research conducted until now.

5.1 Summary

In this dissertation, the following approaches have been proposed related to word-level QE in translation technologies:

- word-level QE in TM-based CAT based on word alignments **[Section 2.1, Chapter 2]**;
- word-level QE in TM-based CAT based on any SBI **[Section 2.2, Chapter 2]**; and
- word-level QE in MT based on any SBI **[Chapter 3]**.

The objective of these methods is to improve the productivity of translators when using these translation technologies by helping them to easily detect which parts of the translation hypotheses need post-editing. In addition to these methods, this dissertation analyses a strategy to provide additional SBI to enable the application of the word-level QE methods using SBI to under-resourced language pairs:

- using the tool Bitextor to crawl parallel data from the Internet **[Chapter 4]**.

The main working hypothesis guiding this research has been that **it is possible to develop methods exclusively based on external SBI for word-level QE in TM-based CAT and MT**. Table 5.1 lists all the partial working hypotheses derived from it and summarises the main contributions to the state of the art obtained as a result of their confirmation.

The approaches described in Chapter 2 are, to the best of our knowledge, the first methods proposed for word-level QE in TM-based CAT. The lack of previous works on word-level QE in TM-based CAT motivated the experiments with professional translators conducted in Appendix A of reprinted publication 2.2.1 to prove the relevance of the problem. The results obtained confirm that providing professional translators with an estimation of the quality of the translation suggestions at the word level has a considerable impact in their productivity: up to about 14% of time can be saved in a translation task.

Two families of methods were defined for word-level QE in TM-based CAT: one using word alignments, and another one directly using SBI. The methods directly using SBI obtained better results and proved to be more robust to domain heterogeneity. This family of methods can in turn be divided in two different approaches for word-level QE in TM-based CAT: one heuristic and another based on binary classification. The heuristic approach has been implemented as a plugin for the free/open-source TM-based CAT tool OmegaT: quality estimations are shown by colouring words in the translation proposals. This approach for word-level QE in TM-based CAT is the only one that does not need to be trained, which brings about two additional advantages: (i) it can be used *on the fly* for any new translation tasks; and (ii) it is not only robust to domain heterogeneity, but also to the language pairs used in the translation process, because it does not need to be trained and makes no assumptions about the languages involved.

The work described in Chapter 2 also allowed obtaining new word-alignment methods based on SBI as a by-product. Namely, two word-alignment approaches were proposed: an heuristic approach that does not need to be trained (see reprinted publication 2.1.1), and a more general maximum-likelihood approach (see reprinted publication A.1). These are the first methods capable of using any SBI in an agnostic way in order to obtain word alignments, that is, they make no assumptions regarding the amount, quality or source of the bilingual information used. These approaches are able to produce word alignments with a precision comparable to that obtained by the state-of-the-art system GIZA++ (Och and Ney, 2003), which is based on statistical word-alignment models. Even though the recall of the methods using SBI is lower than that obtained by GIZA++, they can be still useful for tasks where precision is more important than accuracy. In addition, when having small amounts of parallel data to train statistical word-alignment models, methods based on SBI perform better in general.

A novel method for word-level QE in MT that uses SBI was defined in Chapter 3. The method extracts positive and negative features and uses a *multilayer-perceptron* binary classifier to decide if a word needs to be post-edited or not. To the best of our knowledge, this is the first approach that do not rely on an specific source of

Working hypothesis	Contributions	Reference
Hypothesis #1: it is possible to use word alignments to estimate the quality of TM-based CAT suggestions at the word level.	✓Novel method for word-level QE in TM-based CAT ✓Application of statistical word alignment to word-level QE in TM-based CAT	Reprinted publication 2.1.1
Hypothesis #2: it is possible to use any SBI to obtain word alignments.	✓SBI-based word-alignment method as precise as state-of-the-art statistical word alignment methods	Reprinted publications 2.1.2 and A.1
	✓Training-free word alignment with an heuristic SBI-based method	Reprinted publication 2.1.2
Hypothesis #3: it is possible to use SBI to estimate the quality of TM-based CAT translation suggestions at the word level.	✓Best performing word-level QE method in TM-based CAT ✓Training-free word-level QE in CAT	Reprinted publication 2.2.1
Hypothesis #4: it is possible to take the SBI-based methods for word-level QE in TM-based CAT and adapt them for their use in MT.	✓Word-level QE in MT method non-dependent of any specific source of information	Reprinted publication 3.1
	✓Best system in QE shared task at the WMT 2015	Reprinted publication 3.2
Hypothesis #5: it is possible to create new SBI that enable word-level QE for language pairs with no SBI available using Bitextor to crawl parallel data.	✓Intrinsic evaluation of Bitextor for building English–Croatian parallel corpora	Reprinted publication 4.1
	✓Extrinsic evaluation of Bitextor for building English–Croatian phrase-based SMT system	Reprinted publication 4.2
Hypothesis #6: the results obtained in word-level QE for under-resourced language pairs can be improved by using new SBI obtained through parallel data crawling.	✓Extrinsic evaluation of Bitextor for word-level QE in TM-based CAT for an under-resourced language pair ✓Evaluation of SBI created *ad-hoc* for SBI-based word-level QE in TM-based CAT	Appendix B.2

Table 5.1: Table summarising the partial working hypothesis in this dissertation and relating them to the contributions to the state of the art in word-level QE of translations.

information and it is, therefore, able to use any SBI in an agnostic fashion. This new method for word-level QE is described and evaluated in reprinted publications 3.1 and 3.2. For the evaluation, we used the datasets provided by the organisers of the shared task for word-level QE in MT in the Workshop on Statistical Machine Translation, in the 2014 and 2015 editions, respectively. The approach proposed performed surprisingly well when compared to other approaches participating in the workshop: it obtained results close to those obtained by the best performing approaches in the 2014 edition, while it outperformed the rest of approaches in the 2015 edition. It is worth mentioning that the new approach proposed in Chapter 3 uses only 70 features, far less than those used by other approaches for QE in MT that use binary classification (Camargo de Souza et al., 2014; Biçici and Way, 2014). This makes our approach lighter, specially as regards training.

Chapter 4 explored the use of parallel data crawling to create new sources of bilingual information. In this chapter, the new version 4.1 of the free/open-source tool Bitextor (Esplà-Gomis and Forcada, 2010) was evaluated in the task of obtaining a specific SBI, MT, for an under-resourced language pair: English–Croatian. The tool was extrinsically and intrinsically evaluated; it was first used to build an English–Croatian parallel corpus that was manually evaluated (see reprinted publication 4.1), and then, this corpus was used to build a phrase-based SMT that was evaluated using automatic metrics (see reprinted publication 4.2). Bitextor was compared to another state-of-the-art parallel crawler, the ILSP Focused Crawler (Papavassiliou et al., 2013). Both tools obtained successful results in both evaluations.

Appendix B.2 is aimed at confirming a new use of Bitextor: improving the word-level QE methods in TM-based CAT described in Chapter 2. To do so, some of the experiments in reprinted publication 2.2.1 were repeated including new SBI built with Bitextor for the least-resourced language pair in the experimental setting: English–Finnish. The results obtained show a dramatic improvement of the coverage when combining the SBI used in the original experiments with the new SBI created with Bitextor. These results confirm the usefulness of Bitextor to obtain new SBI on demand for boosting the rest of methods developed in this PhD work.

Finally, it is worth mentioning that this PhD thesis defines, for the first time, strategies to estimate the quality of translation suggestions both for MT and TM-based CAT using any SBI. This means that both facilities could be implemented in a single CAT environment to estimate the quality for both technologies using the same SBI collection. In this way, the translator would obtain quality estimations for both technologies in a transparent way, that is, without having to provide specific information or models to estimate the quality for TM-based CAT and for TM separately. It could be even possible to integrate Bitextor in this tool as well; this would allow translators to create their own SBI for a specific translation task by only providing a collection of URLs of multilingual websites.

5.2 Future work

The research described in this dissertation open up new lines of research. Some of the most evident new lines of work are:

- *a better analysis of how professional translators can use word-level QE*: the experiments on word-level QE in TM-based CAT conducted in reprinted publication 2.2.1 with professional translators raised some questions that had not been previously considered. For example, the fact that a minimum coverage (i.e. percentage of words for which a quality estimation is provided) of word-level QE for a given translation suggestion is needed to make the method useful. In the experiments conducted in this article the coverage of the method was evaluated at the level of the whole translation job. However, it may happen that the coverage is much lower for a specific translation suggestion. In these cases, it may be better not to provide estimations for any word in the translation suggestion. In addition, the errors in the estimation of the quality of some words in TM-based CAT may have different costs: for example, an error in the estimation of a wrong word may be much worse than the cost of an error in the estimation of a correct word. Modelling cost functions and using them to train the QE system could result in higher productivity for translators.

  These are aspects that can be hardly evaluated with automatic experiments and would need to involve professional translators to evaluate their impact in terms of translation productivity. Given that all the methods presented in this dissertation have been automatically evaluated, it is extremely interesting to evaluate them in a professional translation environment, in order to measure their real impact on translation productivity.

- *QE for parallel corpora crawling*: word-level QE is studied in this dissertation in the framework of translation technologies. Therefore, the methods proposed are aimed at improving the experience of using the translations produced by these technologies. However, some of the methods developed may be applied to estimate the quality of parallel texts, and not only those produced by translation technologies. For example, the methods described in Chapter 3 do not make any assumptions about the technology used to produce the translations. It may be interesting to study the application of these technologies in new scenarios, for example, to estimate the quality of parallel corpora. In this way, it would be possible to estimate the general quality of a parallel corpus before using it, for example, to train a SMT system, or to use it in a CAT environment as a TM. The task of estimating the quality of a corpus could be seen as a problem of data selection (Banerjee et al., 2015; Axelrod, 2014) in which the objective is to choose the corpus whit the highest possible quality for a translation task. It could also be possible to use these techniques to remove low-quality TUs, or even to remove parts of TUs that are inadequate translations, keeping those parts which are correctly translated, as done by several authors when trying to extract parallel data from comparable corpora (Zhao and Vogel, 2002; Fung and Cheung, 2004; Munteanu and Marcu, 2006). These techniques could also be integrated in the tool Bitextor to improve its performance to align documents, or to clean the parallel corpus crawled while it is being built.

- *automatic or aided post-editing*: one of the advantages of word-level QE is that it can be helpful for translators when post-editing the output of an MT system or modifying the translation suggestions of a TM-based CAT tool since it can help them to detect the words to be modified. However, it may be possible to go one step further by directly suggesting a translation alternative for a given word to be edited, or even by doing so automatically.

  The word-level QE techniques described in this dissertation are based on using sub-segment pairs (σ, τ) with σ totally matching the source segment S and τ totally or partially matching the translation suggestion T. Negative evidence, that is, evidence that suggest that a word is a wrong translation and therefore should be either replaced or deleted, usually comes from a partial match in which the word receiving this negative evidence could not be matched. This means that the sub-segments τ used to obtain QE may contain translation alternatives. For example, suppose the case of the source segment in Catalan S'=*"L'Associació Europea per a la Traducció Automàtica"*, which could be translated by an MT system as T=*"The European Association for Automatic Translation"*. The word *Automatic* is not the most adequate here, and it should be replaced by *Machine*. A SBI could provide a sub-segment pair (σ, τ)=*("per a la Traducció Automàtica","for Machine Translation")*. The partial match between the TL sub-segment τ=*"for Machine Translation"* and *"for Automatic Translation"* would highlight that *Automatic* is wrong, but the alternative translation *Machine* would not be used by any of the approaches defined in this dissertation as an alternative translation. Therefore, an interesting research line would be to analyse ways of using sub-segment pairs (σ, τ) not only to estimate quality, but also to suggest translation alternatives. Indeed, Ortega et al. (2014) explore the possibility of using SBI for automatically post-editing the translation suggestions of a TM-based CAT tool, a task that they term as *fuzzy-match repair*.

  The strategies mentioned could also be seen as a way to fill a gap that is currently somehow ignored in word-level QE: detecting missing words in translation hypotheses. As already seen in this dissertation, word-level QE focuses on estimating the quality of each of the words occurring in a translation hypothesis. With this information, it is possible to decide which words need to be edited, that is, deleted or replaced. However, insertions, the third possible edition, is not taken into account in most approaches to word-level QE (including those presented in this dissertation). Some work has been done in this direction in a task closely related to this PhD work by using SBI for cross-lingual textual entailment (Esplà-Gomis et al., 2012). In this work, SBI are used to detect if two sentences written in two languages contain the same information. Even though the results obtained in this work were encouraging (it was the second-best performing method in a shared task) the method proposed works at the level of segments and is not yet capable of producing suggestions about which edits need to be done to add new words in a translation hypothesis, which would be definitely much more useful for a translator when post-editing.

- *interactive machine translation*: another interesting extension of the work developed in this PhD thesis is related to IMT (Foster, 2002; Koehn, 2009; Toselli et al., 2011; Torregrosa et al., 2014), also called *interactive translation prediction*. IMT uses MT to guide a professional translator in the process of producing a translation. These programs usually work by interactively suggesting to the professional translator sub-segments[1] that translate the next words of the SL sentence being translated. When the most suitable sub-segment is accepted by the translator, the system keeps suggesting new sub-segments to extend the translation until it is completed; the translator may ignore these suggestions and translate some parts of the SL sentence by hand. Ranking the sub-segments to be presented to the translator and choosing which of them will be shown is a critical problem for these tools (Torregrosa et al., 2014; Sanchis-Trilles et al., 2014). Most IMT systems use information from the inner workings of the MT system used to produce and rank the translation suggestions. Torregrosa et al. (2014), however, proposes a method agnostic as regards the SBI used to obtain the translation suggestions. In such cases, some of the methods proposed in this work could be useful to improve the predictions made to the user. For example, the *on-the-fly* methods for word alignment described in reprinted publication 2.1.2 could be used to detect which words in the SL segment have already been translated and which have not. In the same way, word-level QE in MT methods described in Chapter 3 could be used to discard the translation suggestions containing inadequate words by following the approach by Gandrabur and Foster (2003) and Ueffing and Ney (2005).

[1] These sub-segments may contain from a single word to a complete translation for the whole part of the SL segment that is left to translate.

Appendix A

Extended methods for word alignment using external sources of bilingual information

This appendix describes an extension of the heuristic method proposed in reprinted publication 2.1.2 for a word-alignment method based on SBI. The new method uses a maximum-likelihood-style parametric aligner for which there exists a set of parameter values that makes it equivalent to the heuristic method proposed in reprinted publication 2.1.2. The contents of this appendix correspond to the following technical report:

- Esplà-Gomis, M., Sánchez-Martínez, F., Forcada, M.L. 2012. Using external sources of bilingual information for on-the-fly word alignment, *Technical Report*. Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant. `http://arxiv.org/abs/1212.1192`. **[Reprinted publication A.1]**

The results reported in reprinted publication A.1 confirm that this maximum-likelihood approach outperforms the original heuristic method. The evaluation also compared both SBI-based methods to the state-of-the-art tool GIZA++, which is based on statistical word-alignment models. The results obtained confirmed that when GIZA++ is trained on a large corpus in the same domain to that to be aligned it outperforms the SBI methods described in this appendix. However, when the parallel corpus on which GIZA++ is trained is relatively small (10, 000 pairs of segments or less) or the domain of the training corpus differs from that of the text to be aligned the SBI-based methods perform better. On the contrary, the SBI-based approaches defined in this appendix are robust as regards the domain of the corpora, which makes them more reliable for tasks which require aligning texts in different domains.

The evaluation described in reprinted publication A.1 only focus on the task of word alignment. However, Appendix B.1 describes additional experiments that evaluate them on the problem of word-level QE in TM-based CAT. Figure A.1 highlights the research work reported in this appendix and its relation with the rest of research blocks in this dissertation.

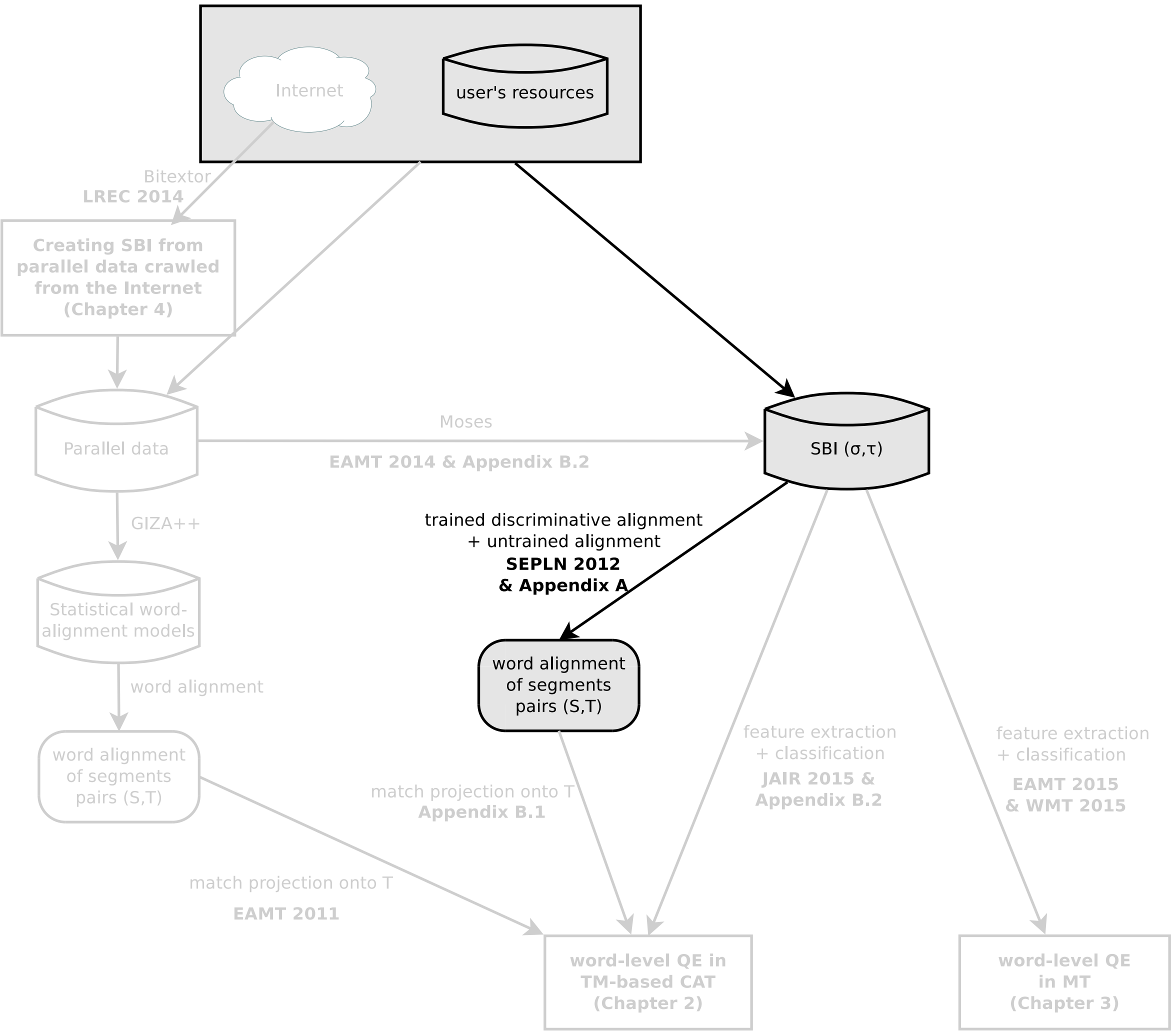

Internet
user's resources
Bitextor
LREC 2014
Creating SBI from parallel data crawled from the Internet (Chapter 4)
Parallel data
Moses
EAMT 2014 & Appendix B.2
SBI (σ,τ)
GIZA++
Statistical word-alignment models
word alignment
word alignment of segments pairs (S,T)
trained discriminative alignment + untrained alignment
SEPLN 2012 & Appendix A
word alignment of segments pairs (S,T)
match projection onto T
Appendix B.1
feature extraction + classification
JAIR 2015 & Appendix B.2
feature extraction + classification
EAMT 2015 & WMT 2015
match projection onto T
EAMT 2011
word-level QE in TM-based CAT (Chapter 2)
word-level QE in MT (Chapter 3)

Figure A.1: This diagram highlights the research being reported in Appendix A (and the publications concerned) on the use of SBI for word alignment. It also shows its relation with the rest of the work reported in this dissertation.

Using external sources of bilingual information for on-the-fly word alignment

Miquel Esplà-Gomis, Felipe Sánchez-Martínez and Mikel L. Forcada
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain
mespla@dlsi.ua.es, fsanchez@dlsi.ua.es, mlf@dlsi.ua.es

Technical report, December 7, 2012

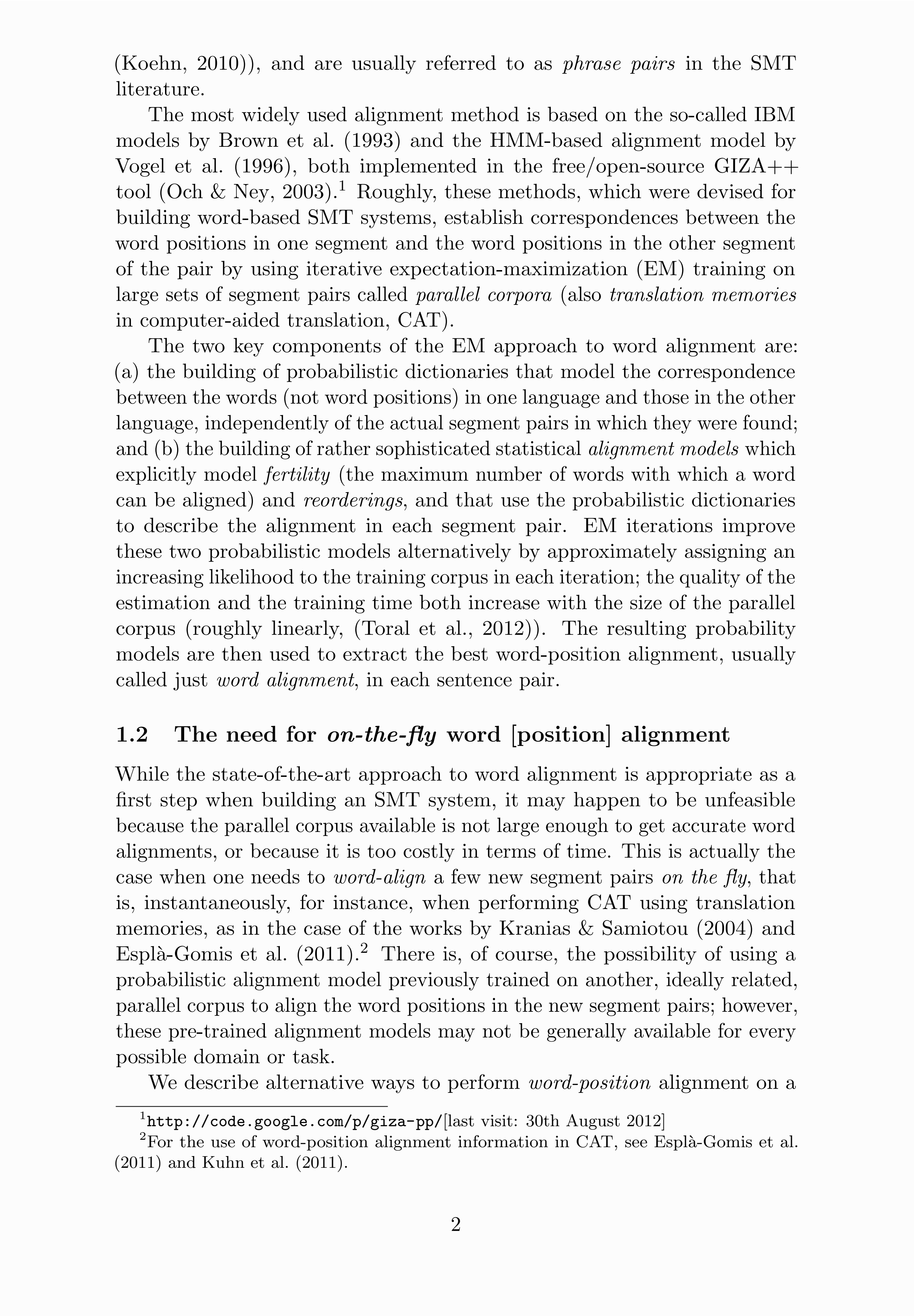
Abstract

In this paper we present a new and simple language-independent method for word-alignment based on the use of external sources of bilingual information such as machine translation systems. We show that the few parameters of the aligner can be trained on a very small corpus, which leads to results comparable to those obtained by the state-of-the-art tool GIZA++ in terms of precision. Regarding other metrics, such as alignment error rate or F-measure, the parametric aligner, when trained on a very small gold-standard (450 pairs of sentences), provides results comparable to those produced by GIZA++ when trained on an in-domain corpus of around 10,000 pairs of sentences. Furthermore, the results obtained indicate that the training is domain-independent, which enables the use of the trained aligner *on the fly* on any new pair of sentences.

1 Introduction

1.1 The need for word [position] alignment

Corpus-based translation technologies use information obtained from existing *segment pairs*, that is, pairs of text segments which are a translation of each other —such as (*Give the book to me*, *Donne-moi le livre*)—, to perform a translation task. These pairs of segments are usually, but not always, *sentence* pairs, and to be able to translate new, unseen text segments, the information in them is usually generalized after performing *word alignment*. The task of word alignment consists in determining the correspondence between the words (actually word positions) in one segment and those in the other segment. After word alignment, smaller sub-segment *translation units*, such as (*le livre*, *the book*), can be extracted. These translation units have a prominent role in state-of-the-art statistical machine translation (SMT,



(Koehn, 2010)), and are usually referred to as *phrase pairs* in the SMT literature.

The most widely used alignment method is based on the so-called IBM models by Brown et al. (1993) and the HMM-based alignment model by Vogel et al. (1996), both implemented in the free/open-source GIZA++ tool (Och & Ney, 2003).[1] Roughly, these methods, which were devised for building word-based SMT systems, establish correspondences between the word positions in one segment and the word positions in the other segment of the pair by using iterative expectation-maximization (EM) training on large sets of segment pairs called *parallel corpora* (also *translation memories* in computer-aided translation, CAT).

The two key components of the EM approach to word alignment are: (a) the building of probabilistic dictionaries that model the correspondence between the words (not word positions) in one language and those in the other language, independently of the actual segment pairs in which they were found; and (b) the building of rather sophisticated statistical *alignment models* which explicitly model *fertility* (the maximum number of words with which a word can be aligned) and *reorderings*, and that use the probabilistic dictionaries to describe the alignment in each segment pair. EM iterations improve these two probabilistic models alternatively by approximately assigning an increasing likelihood to the training corpus in each iteration; the quality of the estimation and the training time both increase with the size of the parallel corpus (roughly linearly, (Toral et al., 2012)). The resulting probability models are then used to extract the best word-position alignment, usually called just *word alignment*, in each sentence pair.

1.2 The need for *on-the-fly* word [position] alignment

While the state-of-the-art approach to word alignment is appropriate as a first step when building an SMT system, it may happen to be unfeasible because the parallel corpus available is not large enough to get accurate word alignments, or because it is too costly in terms of time. This is actually the case when one needs to *word-align* a few new segment pairs *on the fly*, that is, instantaneously, for instance, when performing CAT using translation memories, as in the case of the works by Kranias & Samiotou (2004) and Esplà-Gomis et al. (2011).[2] There is, of course, the possibility of using a probabilistic alignment model previously trained on another, ideally related, parallel corpus to align the word positions in the new segment pairs; however, these pre-trained alignment models may not be generally available for every possible domain or task.

We describe alternative ways to perform *word-position* alignment on a

[1] `http://code.google.com/p/giza-pp/`[last visit: 30th August 2012]

[2] For the use of word-position alignment information in CAT, see Esplà-Gomis et al. (2011) and Kuhn et al. (2011).

segment pair, on the fly and on demand, by using readily available sources of translation units, which we will refer to as *sources of bilingual information* (SBI); for instance, existing (on-line) machine translation systems. Information from the SBI is initially used to discover correspondences between variable-length sub-segments in the pair of segments to align, and then processed to obtain word-position alignments. The word-position alignments are obtained by applying a probabilistic word-position model whose parameters have to be trained on a parallel corpus; no assumptions are made about the pair of languages involved. The corpus, as it will be shown, need not be related to the new segment pairs being word aligned; parameters are therefore transferable across text domains. In addition, there is a particular choice of parameters that completely avoids the need for training and has an intuitive "physical" interpretation, yielding reasonably good results.

1.3 Related work

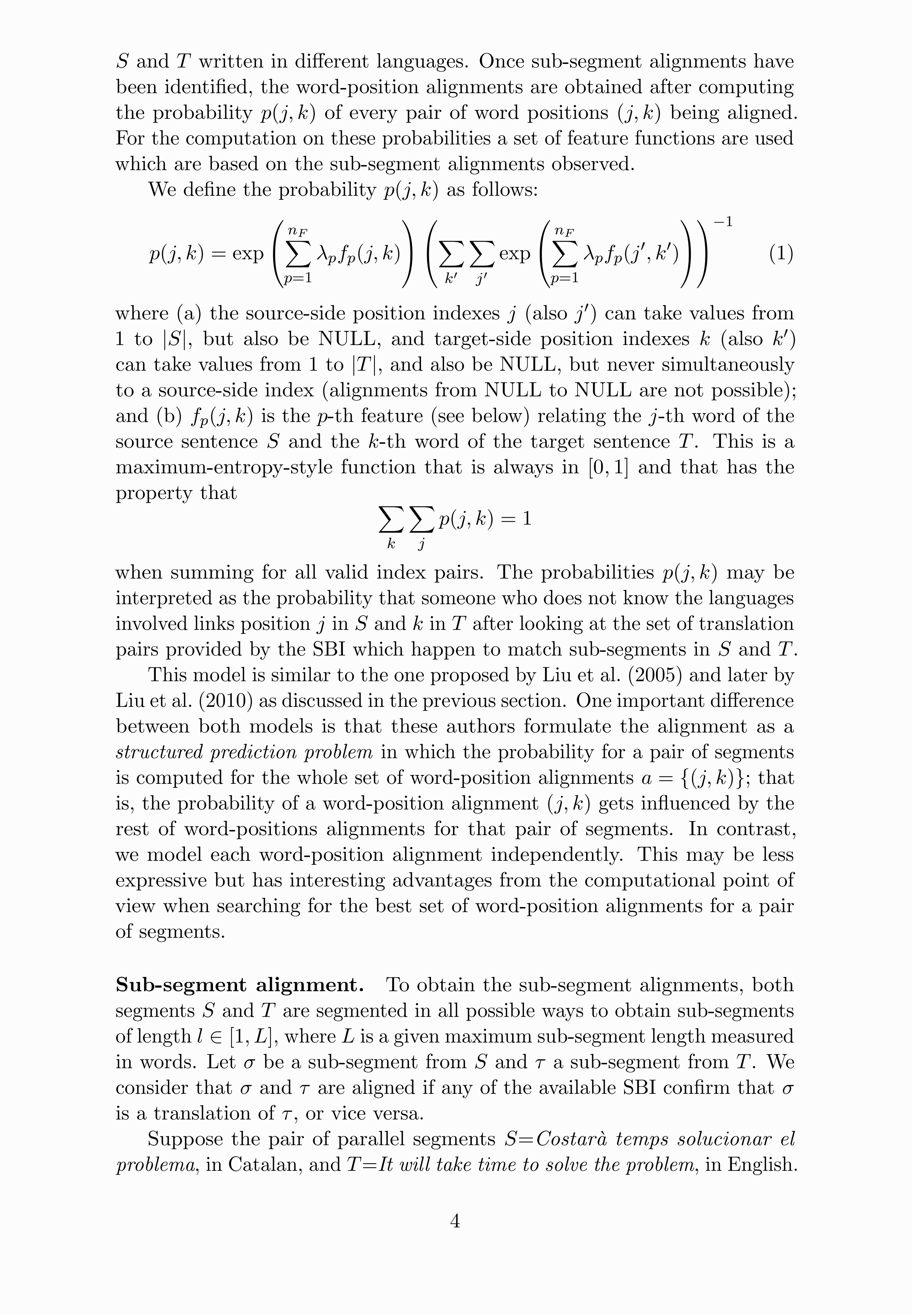
In addition to the IBM models and the HMM alignment model previously mentioned, one can find in the literature different approaches to the problem of word-position alignment. In this section we focus on those approaches that make use of SBI in some way; for a complete review of the state of the art in word alignment the reader is referred to Tiedemann (2011).

Fung & Mckeown (1997) introduces the use of a bilingual dictionary as a SBI to obtain an initial alignment between *seed words* in a parallel corpus. These seed words are chosen so that they cannot have multiple translations (in both languages) and are frequent enough to become useful references in both texts of the parallel corpus. These initial alignments are then used to align the other words appearing around them in the parallel texts using an heuristic method similar to the one introduced by Rapp (1999).

Liu et al. (2005) propose the use of a *log-linear* (maximum-entropy style) model (Berger et al., 1996) to combine the IBM model 3 alignment model with information coming from part-of-speech taggers and bilingual dictionaries; the work was later extended to include new features and a new training procedure (Liu et al., 2010). The main differences between their work and the one presented here are: (i) we do not rely on any previously computed alignment model; (ii) we use any possible SBI which may relate multi-word segments, and (iii) they model the word-position alignment task as a *structured prediction problem* (Tiedemann, 2011, p. 82) that generates the whole alignment structure, whereas we model each association of positions independently. We will further discuss this last difference in the next section.

2 The alignment model

The method we present here uses the available sources of bilingual information (SBI) to detect parallel sub-segments in a given pair of parallel text segments



S and T written in different languages. Once sub-segment alignments have been identified, the word-position alignments are obtained after computing the probability $p(j,k)$ of every pair of word positions (j,k) being aligned. For the computation on these probabilities a set of feature functions are used which are based on the sub-segment alignments observed.

We define the probability $p(j,k)$ as follows:

$$p(j,k) = \exp\left(\sum_{p=1}^{n_F} \lambda_p f_p(j,k)\right)\left(\sum_{k'}\sum_{j'} \exp\left(\sum_{p=1}^{n_F} \lambda_p f_p(j',k')\right)\right)^{-1} \qquad (1)$$

where (a) the source-side position indexes j (also j') can take values from 1 to $|S|$, but also be NULL, and target-side position indexes k (also k') can take values from 1 to $|T|$, and also be NULL, but never simultaneously to a source-side index (alignments from NULL to NULL are not possible); and (b) $f_p(j,k)$ is the p-th feature (see below) relating the j-th word of the source sentence S and the k-th word of the target sentence T. This is a maximum-entropy-style function that is always in $[0,1]$ and that has the property that

$$\sum_k \sum_j p(j,k) = 1$$

when summing for all valid index pairs. The probabilities $p(j,k)$ may be interpreted as the probability that someone who does not know the languages involved links position j in S and k in T after looking at the set of translation pairs provided by the SBI which happen to match sub-segments in S and T.

This model is similar to the one proposed by Liu et al. (2005) and later by Liu et al. (2010) as discussed in the previous section. One important difference between both models is that these authors formulate the alignment as a *structured prediction problem* in which the probability for a pair of segments is computed for the whole set of word-position alignments $a = \{(j,k)\}$; that is, the probability of a word-position alignment (j,k) gets influenced by the rest of word-positions alignments for that pair of segments. In contrast, we model each word-position alignment independently. This may be less expressive but has interesting advantages from the computational point of view when searching for the best set of word-position alignments for a pair of segments.

Sub-segment alignment. To obtain the sub-segment alignments, both segments S and T are segmented in all possible ways to obtain sub-segments of length $l \in [1,L]$, where L is a given maximum sub-segment length measured in words. Let σ be a sub-segment from S and τ a sub-segment from T. We consider that σ and τ are aligned if any of the available SBI confirm that σ is a translation of τ, or vice versa.

Suppose the pair of parallel segments S=*Costarà temps solucionar el problema*, in Catalan, and T=*It will take time to solve the problem*, in English.

We first obtain all the possible sub-segments σ in S and τ in T and then use machine translation as a SBI by translating the sub-segments in both translation directions. We obtain the following set of sub-segment alignments:

temps	↔	*time*
problema	↔	*problem*
solucionar el	→	*solve the*
solucionar el	←	*to solve the*
el problema	↔	*the problem*

It is worth noting that multiple alignments for a sub-segment are possible, as in the case of the sub-segment *solucionar el* which is both aligned with *solve the* and *to solve the*. In those cases, all the sub-segment alignments available are used. Figure 1 shows a graphical representation of these alignments.

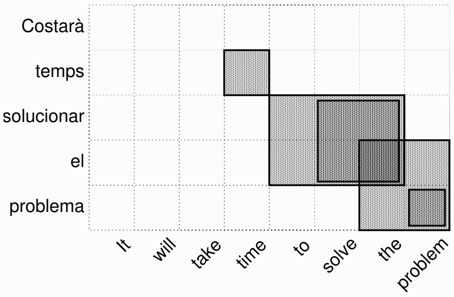


Figure 1: Sub-segment alignments.

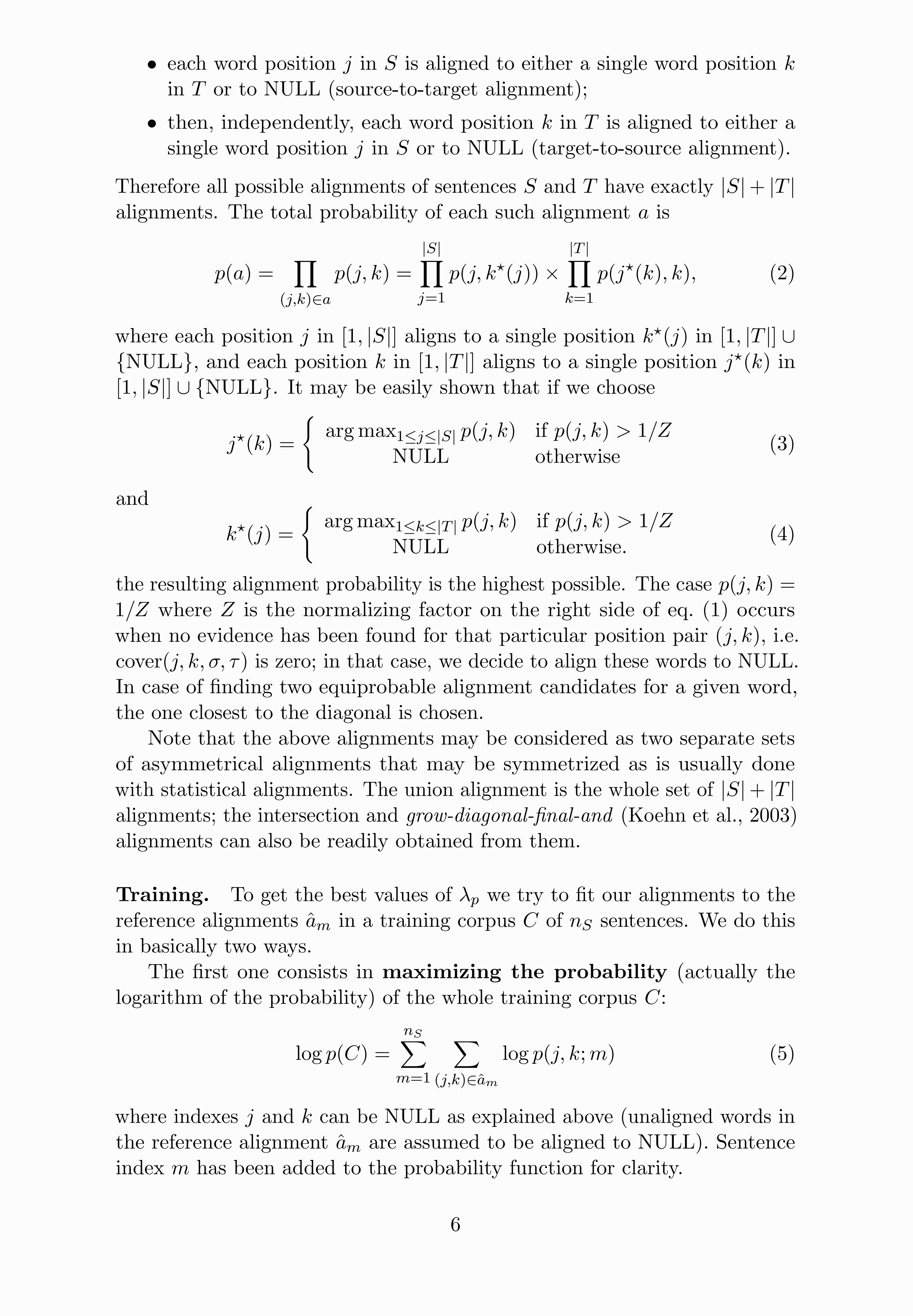
Features. The information provided by the sub-segment alignments is used to build the features that are combined to compute the probabilities $p(j,k)$ through eq. (1). This feature functions are based on the function $\text{cover}(j,k,\sigma,\tau)$, which equals 1 if sub-segment σ *covers* the j-th word in S and τ *covers* the k-th word in T, and 0 otherwise. In particular, by considering sub-segments σ and τ of lengths m and n varying from 1 to the maximum sub-segment length L we define the following set of L^2 features, one feature for each possible combination of lengths $(m,n) \in [1,L] \times [1,L]$:

$$f_{(m-1)L+n} = \sum_{(\sigma,\tau)\in M(S,T),|\sigma|=m,|\tau|=n} \text{cover}(j,k,\sigma,\tau),$$

where $|x|$ stands for the length of sub-segment x measured in words.[3]

Alignment computation. To get the word-position alignments of a pair of segments S and T we follow a greedy method that makes two simplifying assumptions:

[3] One may also split this feature set to treat each different SBI separately or even lift the restriction on the source and target lengths m and n, and build new features depending only on n and m, respectively.



- each word position j in S is aligned to either a single word position k in T or to NULL (source-to-target alignment);
- then, independently, each word position k in T is aligned to either a single word position j in S or to NULL (target-to-source alignment).

Therefore all possible alignments of sentences S and T have exactly $|S|+|T|$ alignments. The total probability of each such alignment a is

$$p(a) = \prod_{(j,k)\in a} p(j,k) = \prod_{j=1}^{|S|} p(j,k^\star(j)) \times \prod_{k=1}^{|T|} p(j^\star(k),k), \tag{2}$$

where each position j in $[1,|S|]$ aligns to a single position $k^\star(j)$ in $[1,|T|] \cup \{\text{NULL}\}$, and each position k in $[1,|T|]$ aligns to a single position $j^\star(k)$ in $[1,|S|] \cup \{\text{NULL}\}$. It may be easily shown that if we choose

$$j^\star(k) = \begin{cases} \arg\max_{1\le j\le |S|} p(j,k) & \text{if } p(j,k) > 1/Z \\ \text{NULL} & \text{otherwise} \end{cases} \tag{3}$$

and

$$k^\star(j) = \begin{cases} \arg\max_{1\le k\le |T|} p(j,k) & \text{if } p(j,k) > 1/Z \\ \text{NULL} & \text{otherwise.} \end{cases} \tag{4}$$

the resulting alignment probability is the highest possible. The case $p(j,k) = 1/Z$ where Z is the normalizing factor on the right side of eq. (1) occurs when no evidence has been found for that particular position pair (j,k), i.e. $\text{cover}(j,k,\sigma,\tau)$ is zero; in that case, we decide to align these words to NULL. In case of finding two equiprobable alignment candidates for a given word, the one closest to the diagonal is chosen.

Note that the above alignments may be considered as two separate sets of asymmetrical alignments that may be symmetrized as is usually done with statistical alignments. The union alignment is the whole set of $|S|+|T|$ alignments; the intersection and *grow-diagonal-final-and* (Koehn et al., 2003) alignments can also be readily obtained from them.

Training. To get the best values of λ_p we try to fit our alignments to the reference alignments $\hat{a}_m$ in a training corpus C of n_S sentences. We do this in basically two ways.

The first one consists in **maximizing the probability** (actually the logarithm of the probability) of the whole training corpus C:

$$\log p(C) = \sum_{m=1}^{n_S} \sum_{(j,k)\in\hat{a}_m} \log p(j,k;m) \tag{5}$$

where indexes j and k can be NULL as explained above (unaligned words in the reference alignment $\hat{a}_m$ are assumed to be aligned to NULL). Sentence index m has been added to the probability function for clarity.

Eq. (5) is differentiable with respect to the parameters λ_p, which allows for gradient ascent training, with each component of the gradient computed as follows:

$$\frac{\partial E}{\partial \lambda_p} = \sum_{m=1}^{n_S} \sum_{(j,k)\in\hat{a}_m} \left(f_p(j,k;m) - \sum_{(j',k')\in\hat{a}_m} p(j',k';m) f_p(j',k';m) \right), \quad (6)$$

where sentence index m has been also added to $f_p(j,k)$ for the sake of clarity.

The second approach tries to **minimize directly an alignment error measure** that indicates how much a discretized, symmetrized alignment obtained by our method departs from the alignments observed in the training corpus: for instance, the alignment error rate (AER) (Och & Ney, 2003) or $1 - F$ where F is the F-measure (Manning & Schütze, 1999, Ch. 8.1), much as it is done by (Liu et al., 2010). Discretization renders these error measures non-differentiable; therefore, we resort to using general-purpose function optimization methods such as the multidimensional simplex optimization of (Nelder & Mead, 1965).[4]

With the two approaches the number of trainable parameters is small (of the order of L^2, where L is the maximum sub-segment length considered), therefore reasonable results may be expected with a rather small training corpus and a SBI covering well the sentence pairs. This is because no word probabilities have to be learned but only parameters to produce word-position alignments using information from the SBIs.

2.1 An intuitive aligner that does not need training

There is a set of parameters for the model described above that has an intuitive “physical” interpretation, and that yields reasonable results, as shown in Section 3. This set of parameters could be used as a starting point for optimization or as a first approximation.

If one chooses $\lambda_{(m-1)L+n} = (mn)^{-1}$, eq. (1) may be rewritten as:

$$p(j,k) = \exp(P_{jk}(S,T,M(S,T))) \left(\sum_{j'} \sum_{k'} \exp(P_{j'k'}(S,T,M(S,T)) \right)^{-1}$$

where the *alignment presssure* $P_{jk}(S,T,M(S,T))$ between the j-th word in S and the k-th word in T is

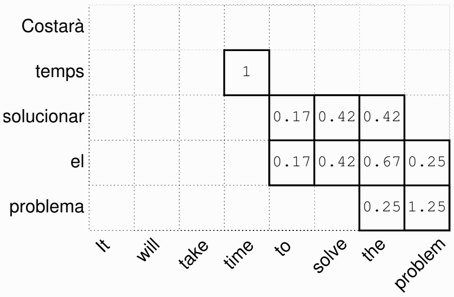
$$P_{jk}(S,T,M(S,T)) = \sum_{(\sigma,\tau)\in M(S,T)} \frac{\mathrm{cover}(j,k,\sigma,\tau)}{|\sigma|\cdot|\tau|}$$

where $M(S,T)$ is the set of sub-segment alignments detected for the pair of parallel segments S and T. If either j or k are NULL, cover(j,k,σ,τ) is zero.

[4] Liu et al. (2010) use MERT instead.

Intuitively, each P_{jk} may be seen as the *pressure* applied by the sub-segment alignments on the word pair (j,k); so the wider the surface $(|\sigma||\tau|)$ covered by a sub-segment alignment, the lower the contribution of that sub-segment pair to the total pressure on (j,k).[5] Clearly, the higher the pressure P_{jk}, the higher the probability $p(j,k)$ is. In the absence of sub-segment information for any of the (j,k)'s of a particular segment pair, all probabilities are equal: $p(j,k) = \frac{1}{(|S||T|+|S|+|T|)}$. The *pressures* are zero when either j or k is NULL.

Following our example, the alignment pressures for the words covered by the sub-segment alignments are presented in Figure 2. The word pair (*temps*,*time*) is only covered by a sub-segment alignment (*temps*, *time*), so the surface is 1 and the alignment pressure is $P_{2,4} = 1$. On the other hand, the word pair (*the*,*el*) is covered by three sub-segment alignments: (*solucionar el*, *solve the*), (*solucionar el*, *to solve the*), and (*el problema*, *the problem*); therefore, the *alignment pressure* is $P_{4,7} = 1/4 + 1/6 + 1/4 = 2/3 \simeq 0.67$.



	It	will	take	time	to	solve	the	problem
Costarà								
temps				1				
solucionar					0.17	0.42	0.42	
el					0.17	0.42	0.67	0.25
problema							0.25	1.25

Figure 2: Alignment pressures.

In this simple model, the alignment pressures P_{jk} themselves may then be used instead of the probabilities $p(j,k)$ to obtain word-position alignments as described at the end of Section 2.

As in the case of the general alignment model defined at the beginning of this section, the alignment is performed both from source-to-target and from target-to-source following the same procedure. Figure 3 shows the Catalan-to-English and the English-to-Catalan word alignments for the running example. As can be seen, words *to* and *solve* in English have the same alignment score for words *solucionar* and *el* in Spanish, respectively. Therefore, the alignments closest to the diagonal are chosen; in this case, *to* is aligned with *solucionar*, and *solve* is aligned with *el* (not a very good alignment). In the other direction of the alignment, the situation is similar for word *solucionar* in Spanish and words *solve* and *the* in English (the resulting alignment is better here).

[5] If just those L^2 features are used and the system is trained on a parallel corpus, the value $mn\lambda_{(m-1)L+n}$ may be considered as the "effective weight" of $m \times n$ sub-segment pairs.

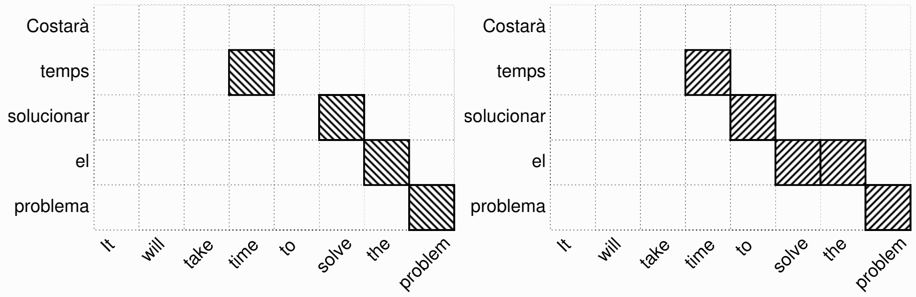


Figure 3: Resulting Catalan-to-English and English-to-Catalan word alignments.

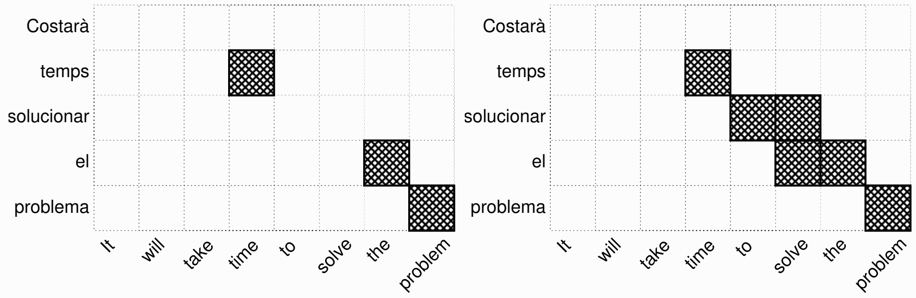


Figure 4: Two possible symmetrized word alignments, the first one using the intersection heuristic and the second one using the *grow-diagonal-final-and* heuristic.

Figure 4 shows two possible symmetrized word alignments obtained by computing, in the first case, the intersection of the alignments shown in figure 3, and, in the second case, the the widely-used *grow-diagonal-final-and* heuristic of Koehn et al. (2003), which, in this case, coincides with the union of the alignments.

3 Experiments

In this section we describe the experimental setting designed for measuring the performance of the alignment models described in Section 2. Two different experimental scenarios were defined in order to measure (a) the quality of the alignments obtained when using training corpora with several levels of reliability, and (b) the domain independence of the weights trained for the parametric aligner (P-aligner).

Gold-standard experiment. For this experiment, we used the EPPS *gold standard* (Lambert et al., 2005), a collection of 500 pairs of sentences extracted from the English–Spanish Europarl parallel corpus (Koehn, 2005) and hand-aligned at the word level using two classes of alignments: *sure* alignments and *possible* alignments.[6] This corpus was used for performing several evaluations:

- *parametric alignment model (defined in Section 2)*: we evaluated this model by using the gold standard corpus both for training and testing using a 10-fold cross-validation strategy. Therefore, for each fold we had 450 pairs of sentences as a training set and 50 pairs of sentences as a test set. We tried the two methods defined in Section 2 for training: optimization of eq. (5) by using a gradient ascent algorithm (Duda et al., 2000), and minimizing directly the alignment error rate (AER) by using the *simplex* algorithm (Nelder & Mead, 1965). Increasingly large sets of bilingual sub-segments were used by defining different values of the maximum sub-segment length L in $[1, 5]$.

- *pressure aligner (defined in Section 2.1)*: Since this alignment model does not require training it was directly evaluated on the gold standard. Increasingly large sets of bilingual sub-segments were used by defining different values of the maximum sub-segment length L in $[1, 5]$.

- *GIZA++ trained on the EPPS gold standard*: GIZA++ (Och & Ney, 2003) was used as a baseline by repeating the previously described 10-fold cross-validation strategy.[7] Although it is obvious that 450 pairs of parallel sentences is not enough for obtaining high quality alignment models with this tool, this results are useful to measure the performance of the models proposed when using a very small training corpus.[8]

- *GIZA++ trained on a large corpus*: In this experiment a larger corpus was used to train GIZA++ models: the English–Spanish parallel corpus provided for the machine translation task at the Seventh Workshop on Statistical Machine Translation (WMT12, Callison-Burch et al. (2012)), which includes the Europarl parallel corpus, from which the gold standard is extracted. In this way, it is possible to compare the models proposed in this work with the use of the state-of-the-art tool GIZA++, which is commonly used in this scenario. This corpus is provided already aligned at the sentence level and, before training the alignment models, it was tokenised and lowercased, and sentences longer than 50 words were removed.[9]

[6] Once the sub-segment alignments were obtained, the gold standard was lowercased to maximise the recall in the alignment process.

[7] The *test-corpus* option in GIZA++ was used to train the alignment models with one corpus and then align another one.

[8] To train GIZA++, the default configuration was used: 5 iterations of model 1 and hidden Markov model and 3 iterations of models 3 and 4.

[9] This preprocessing was performed by using the scripts provided by the Moses MT toolkit: `https://github.com/moses-smt/mosesdecoder/tree/master/scripts`[last visit: 30th August 2012]

corpus en	**05.20.20.10**	**06.30.10.00**
02.40.10.40	0.21	0.18
06.30.10.00	0.15	

corpus es	**05.20.20.10**	**06.30.10.00**
02.40.10.40	0.22	0.18
06.30.10.00	0.13	

Table 1: Cosine similarity for both the English (en) and the Spanish (es) documents in the corpora released by the *European Commission Directorate-General for Translation* that we used.

Since all the alignment models proposed in this experiment are asymmetric (i.e. they must be trained from English to Spanish and from Spanish to English separately) we experimented three different symmetrization methods: intersection, union, and *grow-diagonal-final-and* (Koehn et al., 2005).

GIZA++ alignments as a reference. This second experiment focuses on measuring the re-usability of the weights trained for the parametric alignment model. In this case, we used three different corpora, all of them extracted from the translation memory published by the *European Commission Directorate-General for Translation* (European Commission, 2009).[10] This translation memory is a collection of documents from the *Official Journal of the European Union*[11] which are provided aligned at the sentence level. These documents are indexed by using a set of domain codes[12] which can be used to identify the documents belonging to the same domain. Following this method, we extracted three subsets from this translation memory belonging to the domains: *elimination of barriers to trade* (code 02.40.10.40), *safety at work* (code 05.20.20.10), and *general information of public contracts* (code 06.30.10.00). These corpora were chosen because they have similar sizes (between 15894 and 13414 pairs of sentences) and they belong to clearly different domains, as evidenced by the cosine similarity measure[13] presented in Table 1.[14]

For this experiment, we followed these steps:

- GIZA++ was used to align the three corpora and these alignments

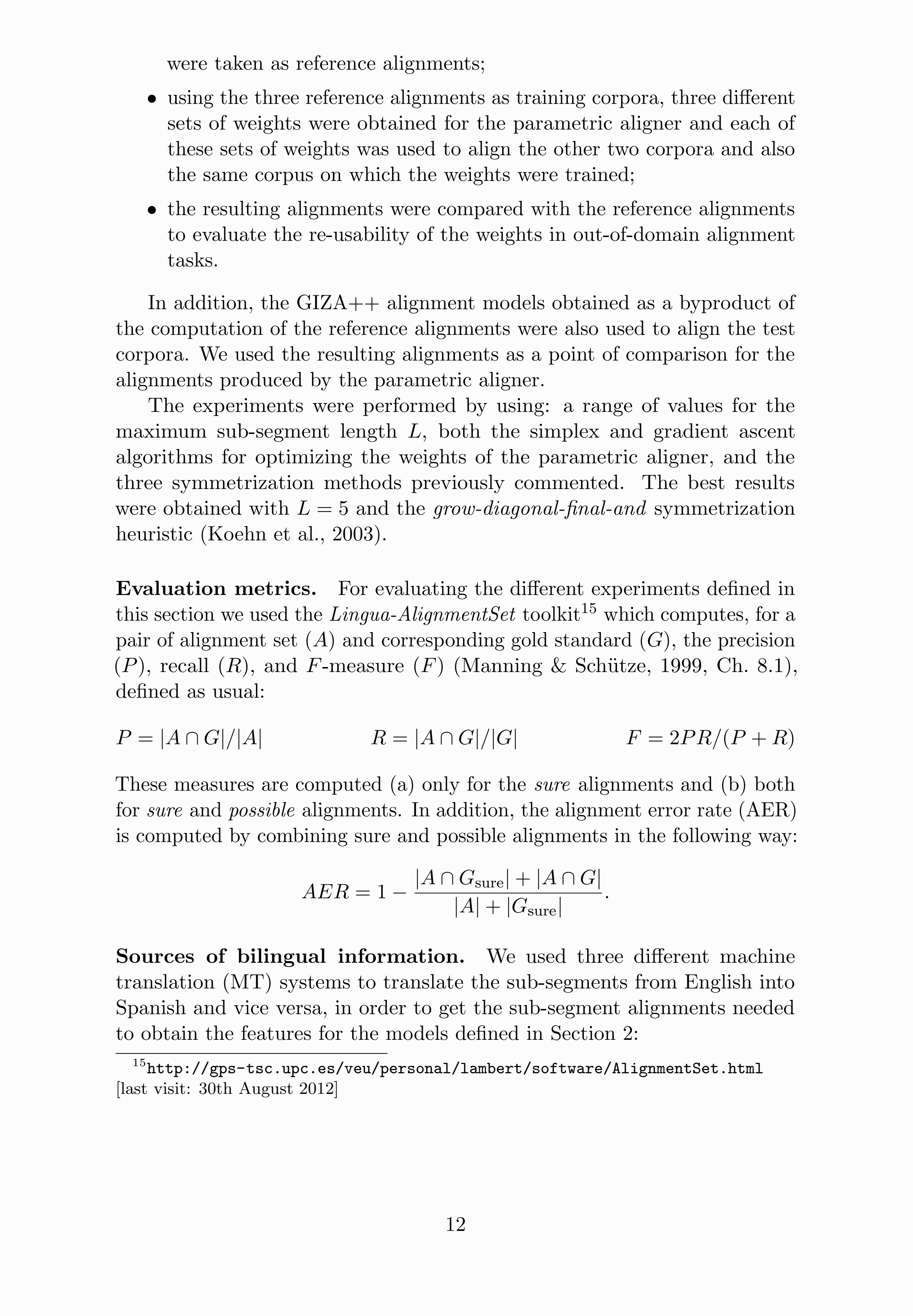
[10] `http://langtech.jrc.it/DGT-TM.html`[last visit: 30th August 2012]

[11] `http://eur-lex.europa.eu`[last visit: 30th August 2012]

[12] `http://eur-lex.europa.eu/RECH_repertoire.do`[last visit: 30th August 2012]

[13] The cosine similarity was computed on the lowercased corpora, removing the punctuation signs and the stopwords defined in the Snowball project: `http://snowball.tartarus.org/algorithms/english/stop.txt`,
`http://snowball.tartarus.org/algorithms/spanish/stop.txt`[last visit: 30th August 2012]

[14] As a reference, note that if we split any of these three corpora into two parts and compute the cosine similarity between them, the results obtained are around 0.98.



were taken as reference alignments;

- using the three reference alignments as training corpora, three different sets of weights were obtained for the parametric aligner and each of these sets of weights was used to align the other two corpora and also the same corpus on which the weights were trained;
- the resulting alignments were compared with the reference alignments to evaluate the re-usability of the weights in out-of-domain alignment tasks.

In addition, the GIZA++ alignment models obtained as a byproduct of the computation of the reference alignments were also used to align the test corpora. We used the resulting alignments as a point of comparison for the alignments produced by the parametric aligner.

The experiments were performed by using: a range of values for the maximum sub-segment length L, both the simplex and gradient ascent algorithms for optimizing the weights of the parametric aligner, and the three symmetrization methods previously commented. The best results were obtained with $L = 5$ and the *grow-diagonal-final-and* symmetrization heuristic (Koehn et al., 2003).

Evaluation metrics. For evaluating the different experiments defined in this section we used the *Lingua-AlignmentSet* toolkit[15] which computes, for a pair of alignment set (A) and corresponding gold standard (G), the precision (P), recall (R), and F-measure (F) (Manning & Schütze, 1999, Ch. 8.1), defined as usual:

$$P = |A \cap G|/|A| \qquad R = |A \cap G|/|G| \qquad F = 2PR/(P + R)$$

These measures are computed (a) only for the *sure* alignments and (b) both for *sure* and *possible* alignments. In addition, the alignment error rate (AER) is computed by combining sure and possible alignments in the following way:

$$AER = 1 - \frac{|A \cap G_{\text{sure}}| + |A \cap G|}{|A| + |G_{\text{sure}}|}.$$

Sources of bilingual information. We used three different machine translation (MT) systems to translate the sub-segments from English into Spanish and vice versa, in order to get the sub-segment alignments needed to obtain the features for the models defined in Section 2:

[15] `http://gps-tsc.upc.es/veu/personal/lambert/software/AlignmentSet.html` [last visit: 30th August 2012]

- *Apertium*:[16] a free/open-source platform for the development of rule-based MT systems (Forcada et al., 2011). We used the English–Spanish MT system from the project's repository[17] (revision 34706).
- *Google Translate*:[18] an online MT system by Google Inc. (translations performed in July 2012).
- *Microsoft Translator*:[19] an online MT system by Microsoft (translations performed in July 2012).

It is worth noting that the Apertium system is oriented to closely-related pairs of languages; furthermore, the Spanish–English language pair is not as mature as other pairs in Apertium; therefore, it is expected to produce translations of lower quality compared with other state-of-the-art systems as indicated by observed BLEU scores. For the gold-standard experiment, these three MT systems were used. For the experiments using the translation memories released by the *European Commission Directorate-General for Translation*, only Apertium and Google could be used, given the huge amount of sub-segments to be translated and the restrictions in the Microsoft Translator API.

4 Results and discussion

This section presents the results obtained in the experiments described in the Section 3. Table 2 shows the results in terms of precision (P), recall (R), F-measure (F) and alignment error rate (AER) obtained by both the parametric aligner (P-aligner) described in Section 2, the "pressure" aligner described in Section 2.1, and GIZA++ both when using a 10-fold cross-validation strategy on the gold standard corpus and when using the corpus from the WMT12 workshop for training the alignment models. It is worth noting that the results computed using the 10-fold cross-validation (*P-aligner probability optimization*, *P-aligner AER optimization*, and *GIZA++ trained on the gold standard*) are presented as the average of the results obtained in each fold. The parametric aligner was both trained by using all the alignments available in the training sets and only using the sure ones. The results of the parametric aligner (best AER in the 27%–29% range) overcame, as expected, the results obtained by the "pressure" aligner (AER around 32%), since the weights were trained on a gold standard and not fixed beforehand.[20] As can be appreciated, both the P-aligner and the "pressure" aligner overcame the results by GIZA++ trained on the gold standard for all the metrics

[16] `http://www.apertium.org` [last visit: 30th August 2012]

[17] `https://apertium.svn.sourceforge.net/svnroot/apertium/trunk/apertium-en-es/` [last visit: 30th August 2012]

[18] `http://translate.google.com` [last visit: 30th August 2012]

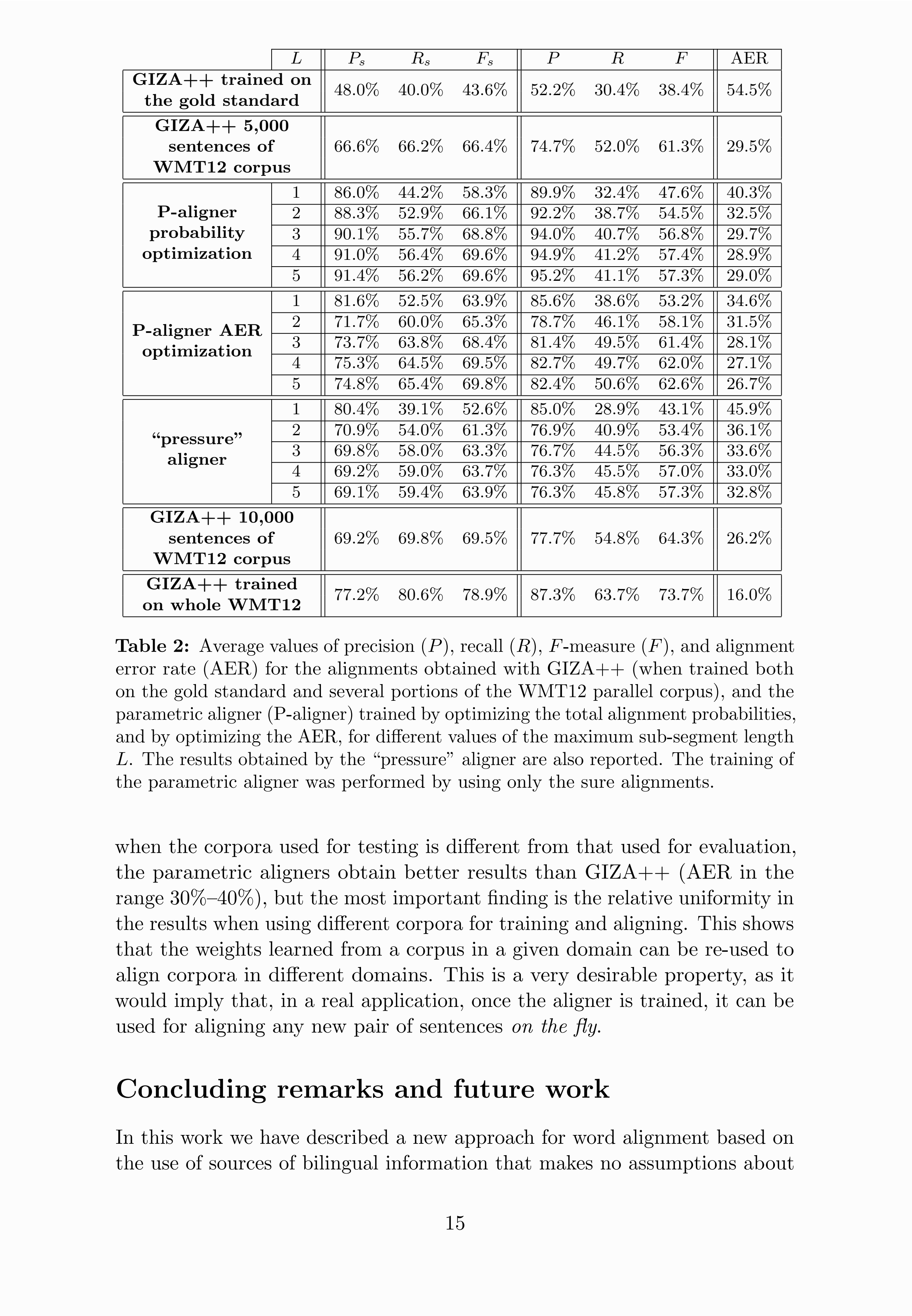
[19] `http://www.microsofttranslator.com` [last visit: 30th August 2012]

[20] The results of the "pressure" aligner come however surprisingly close.

used (AER around 55%). This is easily explainable given the small size of the corpus used to train the alignment models with GIZA++. In any case, this shows the convenience of our model when using a very reduced training corpus. Finally, the alignments from GIZA++ trained on the WMT12 corpus obtained the best results in terms of F-measure and AER (16%). If precision and recall are compared, one can see that the precision in both GIZA++ and the parametric aligner are quite similar but GIZA++ obtains better results in recall. This is an interesting result, since this means that, for tasks like CAT (Esplà-Gomis et al., 2011), where precision is more relevant than the recall, the parametric aligner may be as useful as GIZA++. Also, this means that using more (or better) sources of bilingual information could help to obtain closer results to those obtained by GIZA++ in recall and, consequently, in F-measure and AER. To understand these results better, a complementary experiment was performed by using several sub-sets from the WMT12 corpus with different sizes. We found out that, to obtain the same results produced by the P-aligner in terms of AER, GIZA++ requires an in-domain training corpus with a size between 5,000 pairs of sentences (AER 29.5%) and 10,000 pairs of sentences (AER 26.2%). This confirms that GIZA++ requires a considerably larger training corpus than that needed by the proposed approach and, as a consequence, it would be quite difficult to use it for aligning sentences on the fly or for small amounts of corpora.

There are some differences in the results obtained for the P-aligner depending on the training method used: the model trained through the maximization of the total alignment probability obtained higher results in precision (91% versus 75%), whereas the model trained by minimizing the AER provided better results for recall (65% versus 56%). Although the results for F-measure and AER are very similar, they happen to be slightly better when using the minimization of the AER, as expected, since in this case the evaluation function is directly optimized during the training process.

Finally, Table 3 shows the results obtained for the experiment with the translation memories from the *Official Journal of the European Union*, which is aimed at measuring the domain-independence of the weights trained for the parametric aligner. The table shows, for the parametric aligner (using both training methods) and GIZA++, the results obtained when training on one of the corpora and aligning the other two corpora. The results reported in this table were obtained by using sub-segments of length $L = 5$, as this setting provided the best results. As in the previous experiments, the symmetrization technique used was *grow-diagonal-final-and* (Och & Ney, 2003). As can be seen, the results for all the parametric aligners compared are quite similar for all the systems and all the training/test corpora (AER in the range 27%–34%). It is worth mentioning that in this particular experiment the alignments produced by GIZA++ are being used as a gold standard for evaluation, which could be unfair for our system, since some correct alignments from the P-aligner could be judged as incorrect. Nevertheless,



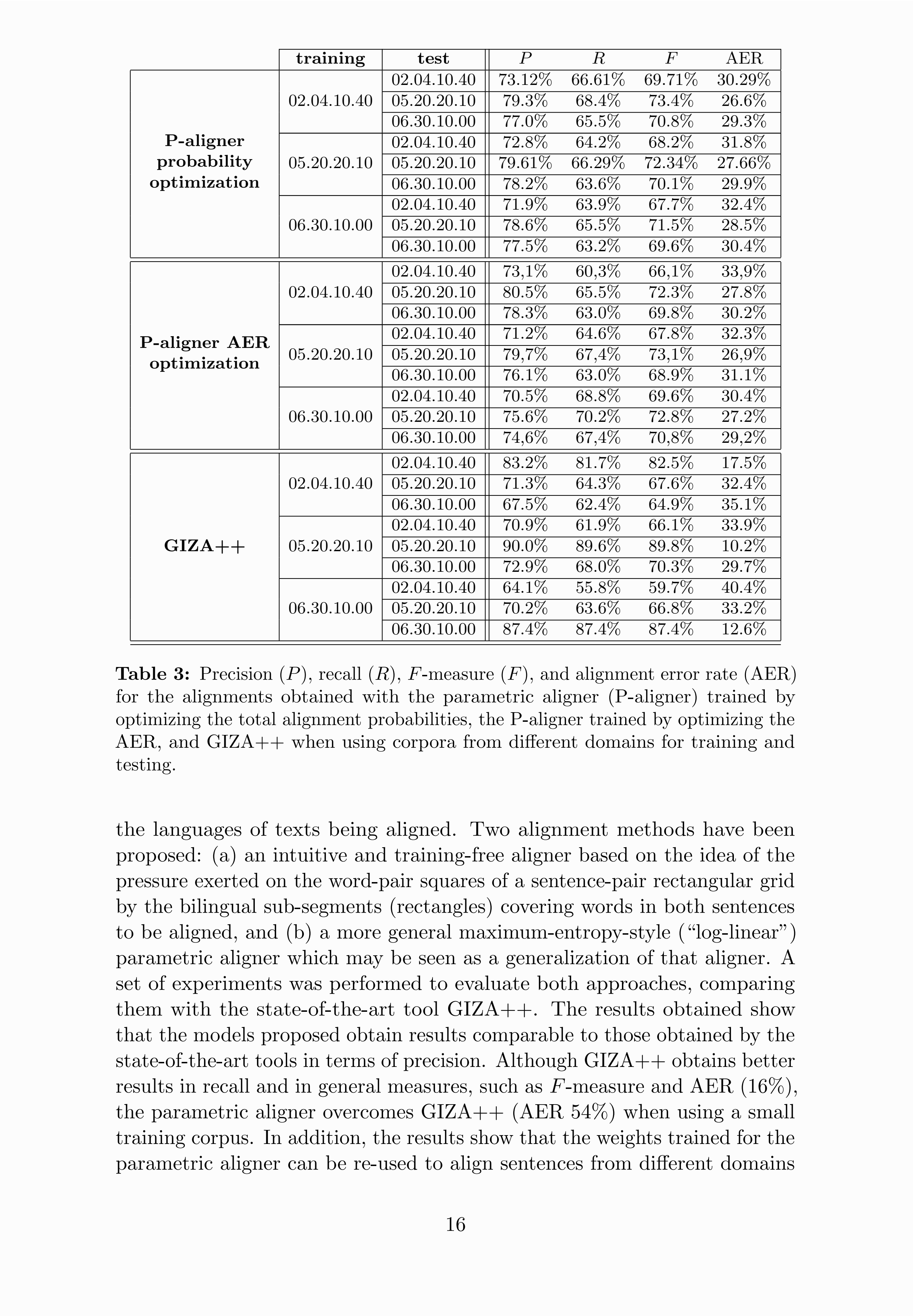
	L	P_s	R_s	F_s	P	R	F	AER
GIZA++ trained on the gold standard		48.0%	40.0%	43.6%	52.2%	30.4%	38.4%	54.5%
GIZA++ 5,000 sentences of WMT12 corpus		66.6%	66.2%	66.4%	74.7%	52.0%	61.3%	29.5%
P-aligner probability optimization	1	86.0%	44.2%	58.3%	89.9%	32.4%	47.6%	40.3%
	2	88.3%	52.9%	66.1%	92.2%	38.7%	54.5%	32.5%
	3	90.1%	55.7%	68.8%	94.0%	40.7%	56.8%	29.7%
	4	91.0%	56.4%	69.6%	94.9%	41.2%	57.4%	28.9%
	5	91.4%	56.2%	69.6%	95.2%	41.1%	57.3%	29.0%
P-aligner AER optimization	1	81.6%	52.5%	63.9%	85.6%	38.6%	53.2%	34.6%
	2	71.7%	60.0%	65.3%	78.7%	46.1%	58.1%	31.5%
	3	73.7%	63.8%	68.4%	81.4%	49.5%	61.4%	28.1%
	4	75.3%	64.5%	69.5%	82.7%	49.7%	62.0%	27.1%
	5	74.8%	65.4%	69.8%	82.4%	50.6%	62.6%	26.7%
“pressure” aligner	1	80.4%	39.1%	52.6%	85.0%	28.9%	43.1%	45.9%
	2	70.9%	54.0%	61.3%	76.9%	40.9%	53.4%	36.1%
	3	69.8%	58.0%	63.3%	76.7%	44.5%	56.3%	33.6%
	4	69.2%	59.0%	63.7%	76.3%	45.5%	57.0%	33.0%
	5	69.1%	59.4%	63.9%	76.3%	45.8%	57.3%	32.8%
GIZA++ 10,000 sentences of WMT12 corpus		69.2%	69.8%	69.5%	77.7%	54.8%	64.3%	26.2%
GIZA++ trained on whole WMT12		77.2%	80.6%	78.9%	87.3%	63.7%	73.7%	16.0%

Table 2: Average values of precision (P), recall (R), F-measure (F), and alignment error rate (AER) for the alignments obtained with GIZA++ (when trained both on the gold standard and several portions of the WMT12 parallel corpus), and the parametric aligner (P-aligner) trained by optimizing the total alignment probabilities, and by optimizing the AER, for different values of the maximum sub-segment length L. The results obtained by the “pressure” aligner are also reported. The training of the parametric aligner was performed by using only the sure alignments.

when the corpora used for testing is different from that used for evaluation, the parametric aligners obtain better results than GIZA++ (AER in the range 30%–40%), but the most important finding is the relative uniformity in the results when using different corpora for training and aligning. This shows that the weights learned from a corpus in a given domain can be re-used to align corpora in different domains. This is a very desirable property, as it would imply that, in a real application, once the aligner is trained, it can be used for aligning any new pair of sentences *on the fly*.

Concluding remarks and future work

In this work we have described a new approach for word alignment based on the use of sources of bilingual information that makes no assumptions about



	training	test	P	R	F	AER
P-aligner probability optimization	02.04.10.40	02.04.10.40	73.12%	66.61%	69.71%	30.29%
		05.20.20.10	79.3%	68.4%	73.4%	26.6%
		06.30.10.00	77.0%	65.5%	70.8%	29.3%
	05.20.20.10	02.04.10.40	72.8%	64.2%	68.2%	31.8%
		05.20.20.10	79.61%	66.29%	72.34%	27.66%
		06.30.10.00	78.2%	63.6%	70.1%	29.9%
	06.30.10.00	02.04.10.40	71.9%	63.9%	67.7%	32.4%
		05.20.20.10	78.6%	65.5%	71.5%	28.5%
		06.30.10.00	77.5%	63.2%	69.6%	30.4%
P-aligner AER optimization	02.04.10.40	02.04.10.40	73,1%	60,3%	66,1%	33,9%
		05.20.20.10	80.5%	65.5%	72.3%	27.8%
		06.30.10.00	78.3%	63.0%	69.8%	30.2%
	05.20.20.10	02.04.10.40	71.2%	64.6%	67.8%	32.3%
		05.20.20.10	79,7%	67,4%	73,1%	26,9%
		06.30.10.00	76.1%	63.0%	68.9%	31.1%
	06.30.10.00	02.04.10.40	70.5%	68.8%	69.6%	30.4%
		05.20.20.10	75.6%	70.2%	72.8%	27.2%
		06.30.10.00	74,6%	67,4%	70,8%	29,2%
GIZA++	02.04.10.40	02.04.10.40	83.2%	81.7%	82.5%	17.5%
		05.20.20.10	71.3%	64.3%	67.6%	32.4%
		06.30.10.00	67.5%	62.4%	64.9%	35.1%
	05.20.20.10	02.04.10.40	70.9%	61.9%	66.1%	33.9%
		05.20.20.10	90.0%	89.6%	89.8%	10.2%
		06.30.10.00	72.9%	68.0%	70.3%	29.7%
	06.30.10.00	02.04.10.40	64.1%	55.8%	59.7%	40.4%
		05.20.20.10	70.2%	63.6%	66.8%	33.2%
		06.30.10.00	87.4%	87.4%	87.4%	12.6%

Table 3: Precision (P), recall (R), F-measure (F), and alignment error rate (AER) for the alignments obtained with the parametric aligner (P-aligner) trained by optimizing the total alignment probabilities, the P-aligner trained by optimizing the AER, and GIZA++ when using corpora from different domains for training and testing.

the languages of texts being aligned. Two alignment methods have been proposed: (a) an intuitive and training-free aligner based on the idea of the pressure exerted on the word-pair squares of a sentence-pair rectangular grid by the bilingual sub-segments (rectangles) covering words in both sentences to be aligned, and (b) a more general maximum-entropy-style ("log-linear") parametric aligner which may be seen as a generalization of that aligner. A set of experiments was performed to evaluate both approaches, comparing them with the state-of-the-art tool GIZA++. The results obtained show that the models proposed obtain results comparable to those obtained by the state-of-the-art tools in terms of precision. Although GIZA++ obtains better results in recall and in general measures, such as F-measure and AER (16%), the parametric aligner overcomes GIZA++ (AER 54%) when using a small training corpus. In addition, the results show that the weights trained for the parametric aligner can be re-used to align sentences from different domains to the one from which they were trained. In this case the new approach provides better results than GIZA++ when aligning out-of-domain corpora. This means that it is possible to use the proposed alignment models to align new sentences *on the fly*, which can be specially useful in some scenarios as the case of computer-aided translation (CAT).

As a future work, we plan to perform wider experiments including other pairs of languages and also other sources of bilingual information. Note that the parameters of the parametric MT-based aligner proposed here could also be intrinsically optimized according to the overall performance of a larger task using alignment as a component, such as *phrase*-based SMT.

References

Berger, A., Della Pietra, V., & Della Pietra, S. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Brown, P., Pietra, S. D., Pietra, V. D., & Mercer, R. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., & Specia, L. (2012). Findings of the 2012 workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

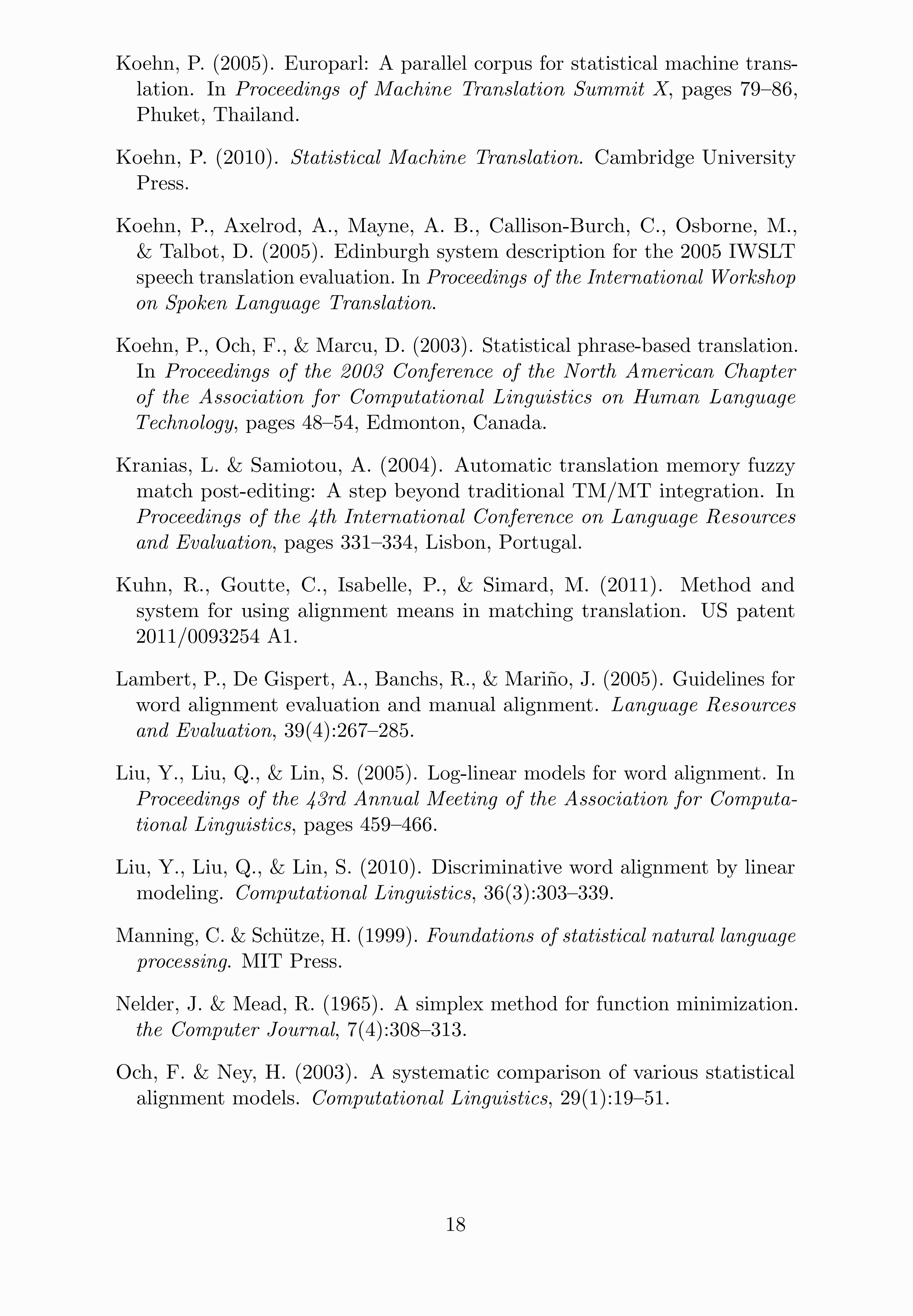
Duda, R., Hart, P., & Stork, D. (2000). *Pattern Classification*. John Wiley and Sons Inc., second edition.

Esplà-Gomis, M., Sánchez-Martínez, F., & Forcada, M. (2011). Using machine translation in computer-aided translation to suggest the target-side words to change. In *Proceedings of the 13th Machine Translation Summit*, pages 172–179, Xiamen, China.

European Commission, D. G. T. (2009). *Translation Tools and Workflow*. Directorate-General for Translation of the European Commission.

Forcada, M., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J., Sánchez-Martínez, F., Ramírez-Sánchez, G., & Tyers, F. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Fung, P. & Mckeown, K. (1997). Finding terminology translations from non-parallel corpora. pages 192–202.



Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, pages 79–86, Phuket, Thailand.

Koehn, P. (2010). *Statistical Machine Translation.* Cambridge University Press.

Koehn, P., Axelrod, A., Mayne, A. B., Callison-Burch, C., Osborne, M., & Talbot, D. (2005). Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation.*

Koehn, P., Och, F., & Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Edmonton, Canada.

Kranias, L. & Samiotou, A. (2004). Automatic translation memory fuzzy match post-editing: A step beyond traditional TM/MT integration. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 331–334, Lisbon, Portugal.

Kuhn, R., Goutte, C., Isabelle, P., & Simard, M. (2011). Method and system for using alignment means in matching translation. US patent 2011/0093254 A1.

Lambert, P., De Gispert, A., Banchs, R., & Mariño, J. (2005). Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation*, 39(4):267–285.

Liu, Y., Liu, Q., & Lin, S. (2005). Log-linear models for word alignment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 459–466.

Liu, Y., Liu, Q., & Lin, S. (2010). Discriminative word alignment by linear modeling. *Computational Linguistics*, 36(3):303–339.

Manning, C. & Schütze, H. (1999). *Foundations of statistical natural language processing.* MIT Press.

Nelder, J. & Mead, R. (1965). A simplex method for function minimization. *the Computer Journal*, 7(4):308–313.

Och, F. & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526, College Park, USA.

Tiedemann, J. (2011). *Bitext Alignment.* Morgan & Claypool Publishers.

Toral, A., Poch, M., Pecina, P., & Thurmair, G. (2012). Efficiency-based evaluation of aligners for industrial apoplications. In Cettolo, M., Federico, M., Specia, L., & Way, A., editors, *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 57–60. Trento, Italy, May 28–30 2012.

Vogel, S., Ney, H., & Tillmann, C. (1996). HMM-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 836–841, Copenhagen, Denmark.

Appendix B

Additional experiments

The objective of this appendix is to report additional experiments that have not been published in peer-reviewed conferences or journals and which help to connect the content in reprinted publications 2.1.2, 4.1, 4.2, and A.1. This appendix consists of two sections:

- Section B.1 revisits the methods to obtain word alignments from SBI described in reprinted publications 2.1.2 and A.1 and includes experiments aimed at evaluating their performance for word-level QE in TM-based CAT; for these experiments the English–Spanish TM described in reprinted publication 2.2.1 is used. The quality of the translation suggestions is estimated at the word level by using the method based on word alignments described in reprinted publication 2.1.1, but using SBI instead of statistical word-alignment models to obtain the word alignments. The results of these experiments are compared to those originally included in reprinted publication 2.2.1.

- Section B.2 includes new experiments for the English–Finnish language pair as regards word-level QE in TM-based CAT. In the experiments described in reprinted publication 2.2.1 English–Finnish was the language pair for which less SBI were available because of the low coverage of resources for Finnish. In this section, some of the experiments described in reprinted publication 2.2.1 are repeated when new SBI crawled by means of the tool Bitextor are added. These experiments are aimed at confirming the usefulness of Bitextor as a support tool for the rest of methods developed in this PhD thesis.

B.1 Using word-alignment methods based on sources of bilingual information for word-level quality estimation in computer-aided translation

Reprinted publications 2.1.2 and A.1 describe new methods to obtain word alignments based on external SBI. The objective of this section is to evaluate the usefulness of these word-alignment methods for the task of word-level QE in TM-based CAT. Figure B.1 highlights the research work covered in this section and its relation with the rest of the work reported in this dissertation.

The methods described in reprinted publications 2.1.2 and A.1 were used to obtain word alignments for the English–Spanish TM in the domain 02.40.10.40 described in Section 6 of reprinted publication 2.2.1 and using the same three SBI: MT systems Apertium, Google Translate and Power Translator. Word-level QE was then obtained by applying the method described in Section 2 of reprinted publication 2.1.1.

Three word-alignment methods based on SBI were evaluated in these experiments:

1. The "pressure" aligner; i.e the heuristic method described in reprinted publication 2.1.2 for word alignment that does not require training.
2. A maximum-likelihood word aligner trained to optimise precision. The word-alignment approach described in reprinted publication A.1 was trained to maximise the alignment precision using a gradient descent algorithm (Duda et al., 2000).
3. A maximum-likelihood word aligner trained for alignment error rate (AER) optimisation: the word-alignment approach described in Section 2 of reprinted publication A.1 was trained to minimise the AER using a multidimensional simplex algorithm (Nelder and Mead, 1965); in this case, gradient descent cannot be used since AER is a discrete metric.

For the three experiments, the *grow-diag-final-and* (Koehn et al., 2005) word-alignment symmetrisation heuristic and the *unanimity* voting criterion were used, since this is the best configuration according to the results obtained in Section 7 of reprinted publication 2.2.1. In addition, for the last two experiments, which require a training process, the English–Spanish EPPS word-alignment gold standard (Lambert et al., 2005) was used for training, as in reprinted publication A.1.

The results of these new experiments were compared to those obtained with the best performing systems for word-level QE using SBI directly and statistical word

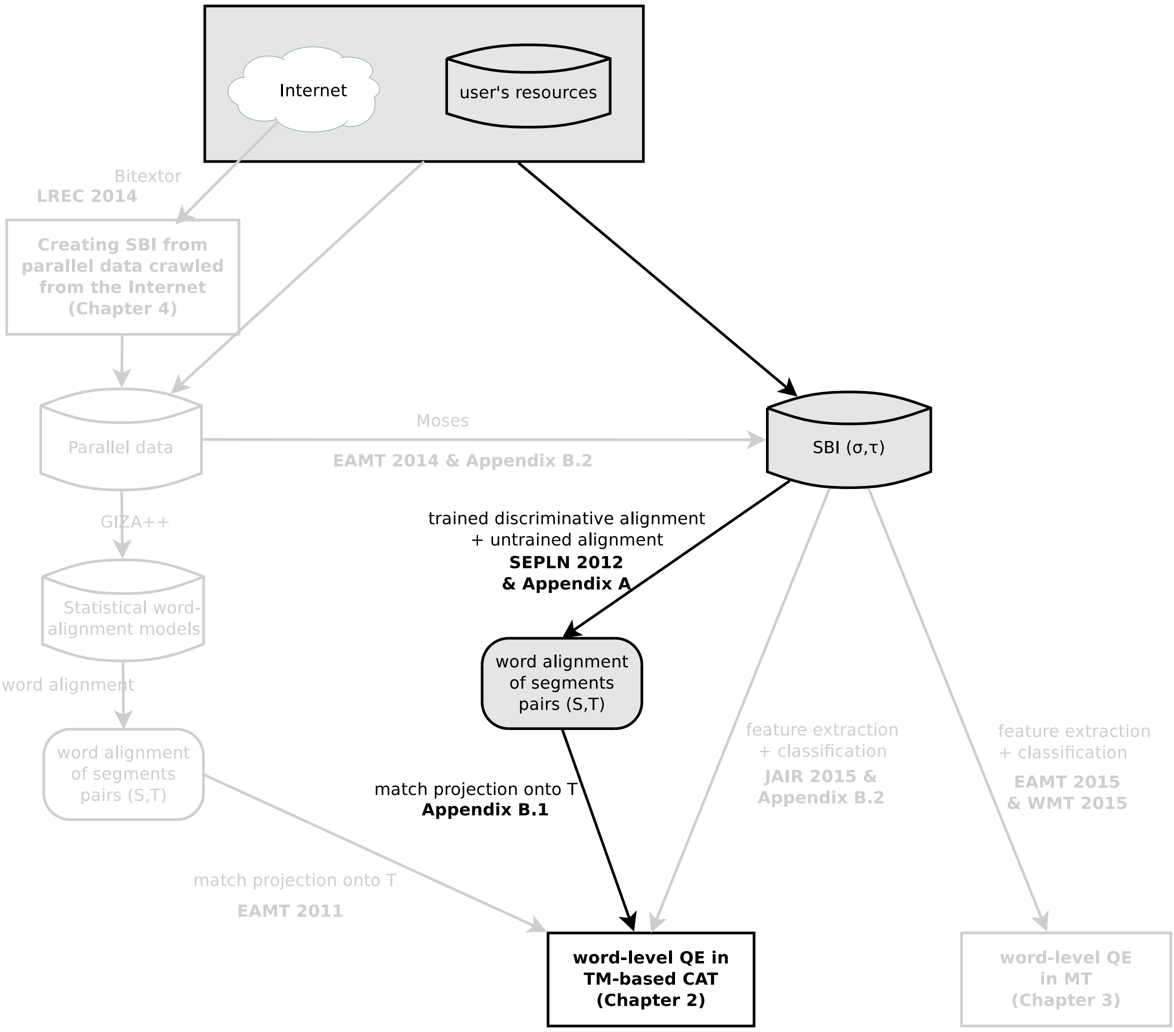

Internet
user's resources
Bitextor
LREC 2014
Creating SBI from parallel data crawled from the Internet (Chapter 4)
Parallel data
Moses
EAMT 2014 & Appendix B.2
SBI (σ,τ)
GIZA++
Statistical word-alignment models
word alignment
word alignment of segments pairs (S,T)
trained discriminative alignment + untrained alignment
SEPLN 2012 & Appendix A
word alignment of segments pairs (S,T)
feature extraction + classification
JAIR 2015 & Appendix B.2
feature extraction + classification
EAMT 2015 & WMT 2015
match projection onto T
Appendix B.1
match projection onto T
EAMT 2011
word-level QE in TM-based CAT (Chapter 2)
word-level QE in MT (Chapter 3)

Figure B.1: This diagram highlights the research being reported in Section B.1 on the use of SBI-based word alignment for word-level QE in TM-based CAT. It also shows its relation with the rest of the work reported in this dissertation.

method	metric	$\Theta \geq 60\%$	$\Theta \geq 70\%$	$\Theta \geq 80\%$	$\Theta \geq 90\%$
pressure	$A(\%)$	94.0±.2	94.7±.2	95.6±.2	96.3±.2
	NC(%)	10.2±.2	9.8±.2	9.4±.3	8.5±.3
precision optimisation	$A(\%)$	94.2±.2	94.9±.2	95.8±.2	96.5±.2
	NC(%)	11.4±.2	11.1±.2	10.7±.3	9.8±.3
AER optimisation	$A(\%)$	94.4±.2	94.9±.2	95.8±.2	96.6±.2
	NC(%)	12.3±.2	11.8±.3	11.3±.3	10.4±.4
statistical word alignment	$A(\%)$	93.9±.2	94.3±.2	95.1±.2	95.3±.3
	NC(%)	6.1±.2	5.9±.2	5.4±.2	**4.9±.3**
PM-C+C	$A(\%)$	**95.1±.1**	**95.6±.2**	**96.4±.2**	**96.9±.2**
	NC(%)	**5.1±.1**	**5.2±.2**	5.5±.2	5.9±.3

Table B.1: For the different values of Θ, accuracy (A) and fraction of words not covered (NC) for word-level QE in TM-based CAT when translating Spanish into English using the corresponding TM from the domain 02.40.10.40 described in Section 6 of reprinted publication 2.2.1. The first three rows show the results obtained when using the method described in Section 2 of reprinted publication 2.1.1, but replacing the statistical word-alignment methods by three SBI-based ones: (i) a maximum-likelihood word aligner trained for precision maximisation, (ii) a maximum-likelihood word aligner trained for AER minimisation, and (iii) a "pressure" word aligner. For three methods the *grow-diag-final-and* word-alignment symmetrisation heuristic and the *unanimity* voting criterion were used. Row 4 reports the results obtained by the original approach described in reprinted publication 2.1.1, which uses statistical word alignments, while row 5 reports the results obtained by the best performing method directly using SBI: the binary classifier using the *PM-C+C* feature collection (see Section 4.3 of reprinted publication 2.2.1). The best results are highlighted in bold if the difference with the rest of results is statistically significant with $p \leq 0.05$.

alignments as reported in Section 7 of reprinted publication 2.2.1: the SBI-based binary classifier using the feature collection termed as "PM-C+C" (see Section 4.3 of reprinted publication 2.2.1) and the statistical word-alignment method trained on the same TM in both translation directions and using the *grow-diag-final-and* symmetrisation heuristic and the *unanimity* voting criterion. For a better comparison, the maximum sub-segment length L was set to 4, as in the experiments described in reprinted publication 2.2.1.

Table B.1 shows the results obtained for word-level QE using SBI-based word alignments. As can be seen, the accuracy obtained by these methods is comparable to that obtained by the best performing SBI-based word-keeping recommendation system and the method using statistical word alignments. On the contrary, the proportion of words not covered is much higher. One would expect to have a similar coverage both when using SBI-based word alignments and when using word-level

method	metric	$\Theta \geq 60\%$	$\Theta \geq 70\%$	$\Theta \geq 80\%$	$\Theta \geq 90\%$
pressure	$A(\%)$	93.7±.2	94.5±.2	95.4±.2	96.0±.2
	NC(%)	10.0±.2	8.9±.2	8.2±.3	7.2±.3
precision optimisation	$A(\%)$	94.2±.2	94.8±.2	95.8±.2	96.4±.2
	NC(%)	9.8±.2	8.8±.2	8.1±.3	7.0±.3
AER optimisation	$A(\%)$	93.7±.2	94.5±.2	95.6±.2	96.4±.2
	NC(%)	9.2±.2	8.4±.2	8.0±.3	7.0±.3
statistical word alignment	$A(\%)$	93.9±.2	94.3±.2	95.1±.2	95.3±.3
	NC(%)	6.1±.2	5.9±.2	5.4±.2	**4.9±.3**
PM-C+C	$A(\%)$	**95.1±.1**	**95.6±.2**	**96.4±.2**	**96.9±.2**
	NC(%)	**5.1±.1**	**5.2±.2**	5.5±.2	5.9±.3

Table B.2: For the different values of Θ, accuracy (A) and fraction of words not covered (NC) for word-level QE in TM-based CAT when translating Spanish into English using the corresponding TM from the domain 02.40.10.40 described in Section 6 of reprinted publication 2.2.1. The first three rows show the results obtained when using the method described in Section 2 of reprinted publication 2.1.1, but replacing the statistical word-alignment methods by three SBI-based ones: (i) a maximum-likelihood word aligner trained for precision maximisation, (ii) a maximum-likelihood word aligner trained for AER minimisation, and (iii) a "pressure" word aligner. For three methods the *union* word-alignment symmetrisation heuristic and the *majority* voting criterion were used. Row 4 reports the results obtained by the original approach described in reprinted publication 2.1.1, which uses statistical word alignments, while row 5 reports the results obtained by the best performing method directly using SBI: the binary classifier using the *PM-C+C* feature collection (see Section 4.3 of reprinted publication 2.2.1). The best results are highlighted in bold if the difference with the rest of results is statistically significant with $p \leq 0.05$.

QE directly obtained through SBI. However, there are two factors that affect negatively the coverage of the methods using word alignments obtained by means of SBI:

- the *grow-diag-final-and* symmetrisation heuristic looses some of the alignments that are not consistent in both alignment directions (Koehn et al., 2005); and
- the *unanimity* voting criterion prioritises the accuracy at the expense of a lower coverage, since the quality of those words for which evidence is contradictory is not estimated.

In order to confirm the impact of these two factors, an additional experiment was carried out using the *union* symmetrisation heuristic, which keeps all the word alignments in both directions when symmetrising, and the *majority* voting criterion, which deals better with contradictory evidence for word-level QE in TM-based CAT.

The objective of this experiment is to check the results obtained by using a configuration that prioritises the coverage instead of the accuracy; the results obtained are shown in Table B.2. As expected, this configuration led to a lower accuracy and to a higher coverage, even though the fraction of words not covered is still higher than that obtained by binary classifier that uses SBI directly. This is due to the fact that, even when using the *majority* voting criterion, no quality estimation can be provided for those words in the TL segment that are aligned to the same number of matched and unmatched words in the SL segment.

B.2 Using Bitextor to build new sources of bilingual information for word-level quality estimation in computer-aided translation

This section is aimed at showing the potential of the tool Bitextor (Esplà-Gomis and Forcada, 2010) to create new SBI for under-resourced language pairs in order to enable the techniques for word-level QE described in this dissertation, as shown in Figure B.2. To this end, the language pair in the experiments described in reprinted publication 2.2.1 with the fewest SBI, English–Finnish, was chosen to do additional experiments with parallel data crawled by using Bitextor. For this task we used a modified version of the English–Finnish corpus published by Rubino et al. (2015), which was obtained by automatically crawling the ".fi" top level domain with the tool SpiderLing (Suchomel et al., 2012), and then aligning the parallel data crawled in the websites explored with Bitextor version 4.1. The difference between the corpus described by Rubino et al. (2015) and the one used in this section is that the *1-best-filtered* setting described in Section 2 of reprinted publication 4.1 was used to build the new version of the parallel corpus. This was done to obtain a corpus with the highest quality possible, given that the experiments in Chapter 4 confirm that this is the best performing setting as regards the accuracy in the alignment of the corpus. The resulting new version of the corpus consisted of about 1,700,000 segment pairs obtained by using Bitextor to extract parallel data from the 10,700 websites crawled with by SpiderLing.

Two kinds of SBI were obtained from this English–Finnish parallel corpus by using the SMT toolkit Moses (Koehn et al., 2007): a phrase table and a phrase-based SMT system. There are some differences in the way in which Moses was used to build SBI:

- *phrase tables*: the phrase tables were obtained by training a standard phrase-based SMT system[1] in both translation directions up to the step in which

[1] http://www.statmt.org/Moses/?n=FactoredTraining.FactoredTraining

phrases are obtained and ranked. In this case, the maximum phrase length was set to 4, since this is the maximum sub-segment length used in the experiments described in reprinted publication 2.2.1, which is the reference here.

- *phrase-based SMT systems*: the parallel corpus obtained was used to train two standard phrase-based SMT for both translation directions. To do so, $1,000$ parallel segments out of the $1,700,000$ segment pairs in the parallel corpus were used for tuning and the rest were used for training. The tuning set was built by randomly choosing segment pairs from the corpus with a confidence score[2] higher than 1.[3] The training set was also used to build both the Finnish and the English language models used by the MT systems. Even though this solution is not optimal (lager language models could be obtained through monolingual crawling, given that the availability of monolingual data is usually higher) this decision was made in order to obtain results using exclusively data obtained with the tool Bitextor.

It is worth noting that, even though the phrase tables and the SMT systems were obtained from the same corpus, there are some noticeable differences between them: on the one hand, phrase tables may contain multiple translations for a single sub-segment, which makes them more informative; on the other hand, MT systems may produce new sub-segment translations that are not in the phrase tables by using shorter phrases.

Tables B.3 and B.4 contain the results involving these new SBI on the task of word-level QE in TM-based CAT for the TM belonging to the 02.40.10.40 domain described in Section 6 of reprinted publication 2.2.1, when translating Finnish into English and the other way round. Google Translate is included as a SBI in these tables because it was the only SBI used in the original experiments in reprinted publication 2.2.1. The results reported correspond to the use of the different SBI separately, as well as the use of different combinations of SBI.

As can be seen, Moses and Google Translate obtain very similar results as regards accuracy; in fact, the difference is not statistically significant in most cases. However, Google Translate has a much higher coverage, while Moses was only able to estimate the quality of slightly more than one half of the words in the test set. As expected, using only phrase tables results in worse accuracy but better coverage. This is due to the fact that phrase tables are much noisier than the other two SBI and the information obtained from them is less precise, but they provide multiple sub-segment translations at once and, therefore, they are more informative. None of the

[2] The confidence score is an estimation of the translation quality at the segment-pair level that is provided by the tool Hunalign (Varga et al., 2005), which is integrated in Bitextor to align the segments of those pairs of documents detected to be parallel.

[3] It is usual to choose a high-quality small amount of parallel data for tuning in SMT in order to obtain better translations (Koehn, 2010).

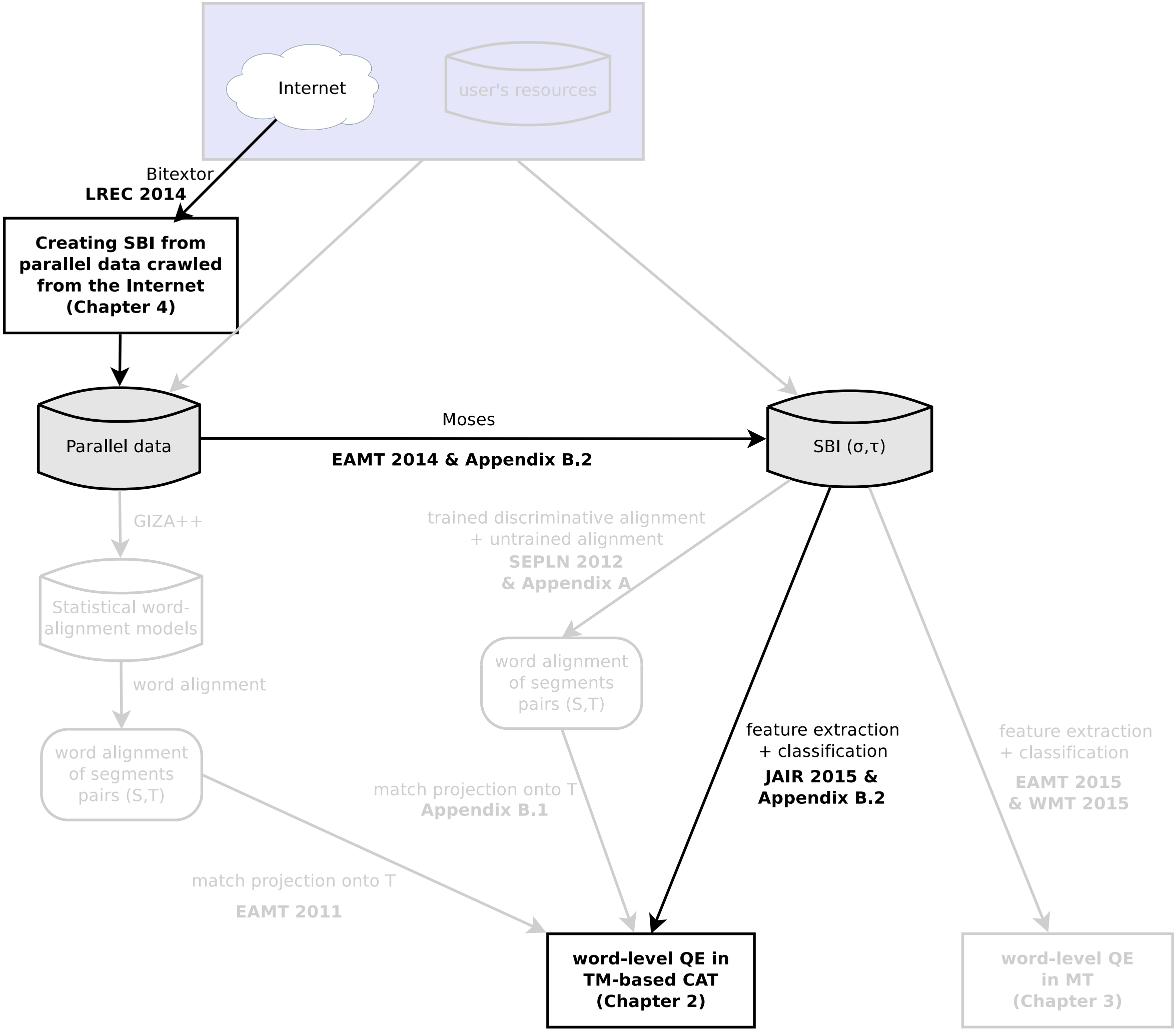

Internet
user's resources
Bitextor
LREC 2014
Creating SBI from parallel data crawled from the Internet (Chapter 4)
Parallel data
Moses
EAMT 2014 & Appendix B.2
SBI (σ,τ)
GIZA++
Statistical word-alignment models
word alignment
word alignment of segments pairs (S,T)
trained discriminative alignment + untrained alignment
SEPLN 2012 & Appendix A
word alignment of segments pairs (S,T)
match projection onto T
Appendix B.1
match projection onto T
EAMT 2011
feature extraction + classification
JAIR 2015 & Appendix B.2
feature extraction + classification
EAMT 2015 & WMT 2015
word-level QE in TM-based CAT (Chapter 2)
word-level QE in MT (Chapter 3)

Figure B.2: This diagram highlights the research being reported in Section B.2 on the impact of SBI built on parallel data crawled with Bitexor on word-level QE in TM-based CAT. It also shows its relation with the rest of the work reported in this dissertation.

new SBI built from parallel data is able to beat Google Translate either in accuracy or in the fraction of words not covered.

It is interesting to observe that when combining Moses and Google Translate, the accuracy remains stable and a small improvement in coverage is obtained of about 10% compared to using only Google Translate. As can be seen, the best coverage is obtained when using all the SBI at the same time: compared to the results obtained with Google Translate, the fraction of words not covered is reduced to less than a half. As expected, a drop in accuracy is observed in this case, which is presumably produced by the noise in the phrase tables. However, this drop is much smaller than the one experienced when using exclusively phrase tables.

The conclusion that can be extracted from these experiments is that parallel data crawling is a successful strategy to build new SBI that can be used to apply SBI-based word-level QE for under-resourced language pairs. It is worth noting that these results are only aimed at proving that Bitextor can be a useful ally when dealing with under-resourced language pairs. However, better approaches could be defined to improve these results. For example, more parallel data could be obtained by combining Bitextor with other state-of-the-art parallel crawlers (Ma and Liberman, 1999; Chen and Nie, 2000; Resnik and Smith, 2003; Désilets et al., 2008; Papavassiliou et al., 2013); as explained in reprinted publication 4.1, combining several parallel data crawlers can result in a richer and larger parallel corpus. Another interesting experiment could be performed by filtering the phrase tables used, for example, using the translation probabilities. This would probably lead to a drop in coverage, but may also improve accuracy. Regarding the Moses phrase-based SMT system, one obvious improvement that could lead to better results is to use a language model trained on larger monolingual corpora; as already mentioned, the availability of monolingual data is usually higher, and it can be easily crawled. For example, using the monolingual corpus described by Rubino et al. (2015) could help to improve the results obtained, especially as regards to coverage. All these ideas exceed the aim of this section, but they deserve to be explored.

Finnish→English						
SBI	**metric**	**method**	$\Theta \geq 60\%$	$\Theta \geq 70\%$	$\Theta \geq 80\%$	$\Theta \geq 90\%$
Google	$A(\%)$	**trained**	93.2±.2	94.6±.2	94.7±.2	94.8±.3
		training-free	91.6±.2	93.4±.3	94.1±.3	94.5±.3
	NC(%)	—	11.2±.2	11.3±.2	11.5±.3	11.6±.4
phrases	$A(\%)$	**trained**	87.7±.3	90.3±.3	91.8±.3	92.3±.4
		training-free	85.9±.3	89.6±.3	91.5±.3	92.2±.4
	NC(%)	—	23.5±.3	23.1±.3	23.1±.4	23.8±.6
Moses	$A(\%)$	**trained**	93.1±.2	94.7±.2	94.9±.3	94.7±.4
		training-free	93.1±.2	94.7±.2	94.9±.3	94.7±.4
	NC(%)	—	45.5±.3	44.2±.4	44.9±.5	46.8±.7
Google	$A(\%)$	**trained**	91.2±.2	92.7±.2	93.1±.3	93.1±.3
+		**training-free**	87.9±.2	90.6±.2	91.9±.3	92.7±.4
phrases	NC(%)	—	5.2±.2	4.8±.2	5.0±.2	5.6±.3
Google	$A(\%)$	**trained**	93.0±.2	94.3±.2	94.4±.2	94.4±.3
+		**training-free**	91.0±.2	93.1±.2	93.7±.3	94.1±.3
Moses	NC(%)	—	10.2±.2	10.2±.3	10.2±.3	9.9±.4
Google	$A(\%)$	**trained**	91.3±.2	92.8±.2	93.2±.3	93.0±.3
+ Moses		**training-free**	87.8±.2	90.5±.2	91.8±.3	92.6±.4
+ phrases	NC(%)	—	5.1±.2	**4.3±.2**	**4.9±.2**	**5.4±.3**

Table B.3: For different FMS thresholds Θ, accuracy (A) and fraction of words not covered (NC) for word-level QE in TM-based CAT when translating Finnish into English using the best trained approach for word-level QE, PM-C+C (see Section 4.3 of reprinted publication 2.2.1), and the training-free approach for word-level QE (see Section 5 of reprinted publication 2.2.1). This table compares both approaches when using three different SBI: Google Translate (Google), the phrase table extracted with Moses from a parallel corpus crawled with Bitextor (phrases), and a fully-trained Moses phrase-based SMT system trained on the same corpus crawled with Bitextor. The results with all the possible combinations of Google and the other two SBI are shown. The word-level QE models for the trained approach were trained on the TM belonging to the 02.40.10.40 domain. The maximum sub-segment length L was set to 4 for all the approaches. Statistically significant differences in the accuracy of each approach for the different values of Θ with $p \leq 0.05$ are highlighted in bold.

English→Finnish						
SBI	**metric**	**method**	$\Theta \geq 60\%$	$\Theta \geq 70\%$	$\Theta \geq 80\%$	$\Theta \geq 90\%$
Google	$A(\%)$	**trained**	89.1±.2	90.2±.3	90.3±.3	90.6±.4
		training-free	86.1±.2	88.6±.3	89.0±.3	90.5±.4
	NC(%)	—	11.6±.2	11.1±.3	12.0±.3	12.6±.4
phrase	$A(\%)$	**trained**	83.3±.3	85.9±.3	96.8±.4	87.4±.4
		training-free	81.0±.3	85.2±.3	86.3±.4	87.3±.4
	NC(%)	—	19.9±.3	20.3±.3	21.1±.4	21.8±.5
Moses	$A(\%)$	**trained**	89.1±.3	**90.8±.3**	**91.1±.3**	**91.1±.5**
		training-free	87.2±.3	90.0±.3	90.5±.3	90.8±.5
	NC(%)	—	31.5±.4	32.0±.4	32.8±.5	34.2±.6
phrase	$A(\%)$	**trained**	86.9±.2	88.1±.3	88.0±.3	98.0±.4
+		**training-free**	82.9±.3	86.1±.3	87.1±.3	87.8±.4
Google	NC(%)	—	4.2±.1	4.0±.2	4.2±.2	5.1±.3
Google	$A(\%)$	**trained**	88.4±.2	89.6±.3	89.7±.3	89.8±.4
+		**training-free**	85.6±.2	88.5±.3	89.1±.3	89.8±.4
Moses	NC(%)	—	8.6±.2	7.9±.2	8.5±.3	9.5±.4
Google	$A(\%)$	**trained**	87.0±.2	88.0±.3	88.1±.3	87.8±.4
+ Moses		**training-free**	82.9±.3	86.0±.3	87.0±.3	87.7±.4
+ phrase	NC(%)	—	**3.9±.1**	**3.7±.2**	**3.9±.2**	**4.7±.3**

Table B.4: For different FMS thresholds Θ, accuracy (A) and fraction of words not covered (NC) for word-level QE in TM-based CAT when translating English into Finnish using the best trained approach for word-level QE, PM-C+C (see Section 4.3 of reprinted publication 2.2.1), and the training-free approach for word-level QE (see Section 5 of reprinted publication 2.2.1). This table compares both approaches when using three different SBI: Google Translate (Google), the phrase table extracted with Moses from a parallel corpus crawled with Bitextor (phrases), and a fully trained Moses phrase-based SMT system trained on the same corpus crawled with Bitextor. The results with all the possible combinations of Google and the other two SBI are shown. The word-level QE models for the trained approach were trained on the TM belonging to the 02.40.10.40 domain. The maximum subsegment length L was set to 4 for all the approaches. Statistically significant differences in the accuracy of each approach for the different values of Θ with $p \leq 0.05$ are highlighted in bold.

Appendix C

Open-source software released as part of this PhD thesis

All of the software developed in the framework of this PhD thesis to evaluate the different methods described has been released under the General Public License version 3. Publishing this software under a free license makes it possible to reproduce the experiments conducted and, in addition, allows anyone to improve it, or use it to develop new methods. This appendix describes each tool and links each software package with the experiments conducted in each chapter.

C.1 `Gamblr-CAT`

`Gamblr-CAT`[1] is a collection of tools that implements the different methods described in reprinted publication 2.2.1 for word-level QE in TM-based CAT. This collection includes three packages, which are implemented in `Java` and are available at `https://github.com/transducens/Gamblr-CAT`. These packages are:

`Gamblr-CAT-alignments` uses word alignments for word-level QE in TM-based CAT (see reprinted publication 2.1.1). It was evaluated through the experiments in reprinted publications 2.1.1, where the tool GIZA++ (Och and Ney, 2003) was used to obtain the word alignments, and in Appendix B.1, where the tool `Flyligner`, described below, was used to obtain word alignments based on SBI.

[1] The name of the package refers to a song by *Kenny Rogers* called *The Gambler*; one of the verses says: *Every gambler knows that the secret to survivin' is knowin' what to throw away and knowin' what to keep.*

Gamblr-CAT-alignments takes as input the SL and TL segments in a TM, together with the alignment between their words in the format used by Moses[2] (Koehn et al., 2007) and the collection of SL segments to be translated. For each segment to be translated, the tool outputs the collection of matching translation units for a given FMS threshold and the word-keeping recommendations for each of them.

It is possible to provide the reference translations for the collection of SL segments to be translated; in this case, the tool performs an evaluation of the word-keeping recommendations produced by checking which words in the translation suggestion remain in the reference translations. This is done by computing a monotonous alignment between the translation suggestions and the references based on the edit distance (Levenshtein, 1966).

Gamblr-CAT-SBI-lib is a library, mainly used by the tool Gamblr-CAT-SBI and the pluggin OmegaT-Marker-Plugin described below. It implements methods to match sub-segments between two segments in different languages and to extract the different collections of features described in Section 4 of reprinted publication 2.2.1 for word-level QE in TM-based CAT. The objective of implementing this as a library is to ease its integration in different tools.

Gamblr-CAT-SBI uses the library Gamblr-CAT-SBI-lib to perform word-level QE in TM-based CAT. This tool was evaluated by means of the experiments described in reprinted publication 2.2.1.

Gamblr-CAT-SBI takes as an input the SL and TL segments of a TM together with a collection of sub-segment pairs, which are the result of splitting the segments in the TM and using SBI to translate them in both translation directions, and the collection of SL segments to be translated. As in Gamblr-CAT-alignments, the tool outputs the collection of matching translation units corresponding to each segment to be translated for a given FMS threshold and the word-keeping recommendations for each of them.

As in Gamblr-CAT-alignments the reference translations can be provided for the collection of SL segments to be translated to evaluate the word-keeping recommendation performance.

C.2 OmegaT-Marker-Plugin

OmegaT-Marker-Plugin is a plugin for the free/open-source TM-based CAT tool OmegaT[3] that implements the heuristic method for word-level QE in TM-based CAT described in reprinted publication 2.2.1. This plugin obtains word keeping

[2] http://www.statmt.org/moses/?n=FactoredTraining.AlignWords

[3] http://www.omegat.org

recommendations by using the on-line MT systems integrated in OmegaT as SBI. It then uses colours to show the recommendations to the user: target words to be kept are coloured in green, while target words to be deleted or replaced are coloured in red.

A modified version of this plugin that used gold-standard recommendations was used to perform the experiments with professional translators described in Appendix A of reprinted publication 2.2.1. It is implemented in `Java` and is available at `https://github.com/transducens/OmegaT-Marker-Plugin`.

C.3 `OmegaT-SessionLog`

`OmegaT-SessionLog` is a plugin for the free/open-source tool for TM-based CAT OmegaT that keeps track of most of the actions performed by the user during translation. This is done in a transparent way, so the user is not disturbed by the plugin. All the actions, as well as the use of any of the translation tools provided by OmegaT (glossaries, MT, translation suggestions from the TM, etc.) are captured by this plugin and stored in an XML file. All the translation actions are logged together with the position of the text in which they were performed and a time stamp. In this way, it is possible to measure the productivity of a translator using the tool, and analyse which is the impact of the different translation tools in this process.

This plugin was used to measure the productivity of professional translators when using word-keeping recommendation in the experiments described in Appendix A of reprinted publication 2.2.1. It is implemented in `Java` and is available at `https://github.com/mespla/OmegaT-SessionLog`.

C.4 `Flyligner`

`Flyligner` is a tool that aligns two segments of parallel text at the word level by means of SBI. This tool is implemented in `Java`, and is available at `https://github.com/transducens/Flyligner`. It implements the heuristic SBI-based method described in reprinted publication 2.1.2, and the maximum-likelihood model, described in reprinted publication A.1. For the method based on maximum-likelihood, two training methods are available: one that maximises the alignment precision by using a gradient descent algorithm, and one that minimises the AER by means of a simplex algorithm.

This software was used to conduct the experiments described in reprinted publication 2.1.2 and in reprinted publication A.1, which show the performance of the new word-alignment methods developed, as well as in Appendix B in which they are applied to TM-based CAT word-level QE.

C.5 Gamblr-MT

Gamblr-MT consists of a collection of scripts in Python 2 that use SBI to extract a collection of features that allow to estimate the quality of MT outputs at the word level. This package is available at https://github.com/transducens/Gamblr-CAT. All the scripts take as input:

- a file containing a tokenised SL sentence per line;
- a file containing a tokenised MT output per line, corresponding to the SL sentences in the first file;
- a file containing a tab-separated list of SL and TL sub-segments; and
- the maximum sub-segment length to be used.

The output of the script is a comma-separated collection of values for each of the features. Every line in the output corresponds to each of the words in the MT outputs evaluated. These scripts implement the features described in Section 3 of reprinted publication 3.1, which are used in the experiments in reprinted publications 3.1 and 3.2.

C.6 Bitextor v.4.1

Bitextor is a tool used to build parallel corpora from multilingual websites. To use it, it is only necessary to provide one or more URLs of websites likely to contain parallel data, as well as a bilingual lexicon for the languages to be crawled. This tool is mainly implemented in Bash and Python 2, and is available at http://www.softonic.net/p/bitextor. A wiki[4] is also available that contains most of the documentation, including an installation guide and a small tutorial.

Bitextor has been developed at Universitat d'Alacant[5] and its first version is previous to the beginning of this PhD thesis. However, it has been significantly improved within the framework of the EU-funded project Abu-MaTran in collaboration with the company *Prompsit Language Engineering*.[6]

[4] http://sourceforge.net/p/bitextor/wiki/Home/

[5] http://www.ua.es

[6] http://www.prompsit.com

In this dissertation, `Bitextor` is used in Chapter 4 to create an English–Croatian parallel corpus (see reprinted publication 4.1) which is then used to train a phrase-based SMT system (see reprinted publication 4.2). `Bitextor` is also used in Appendix B.2 to create new English–Finnish SBI that are evaluated for word-level QE in TM-based CAT.

Index of abbreviations

SL Source language 1

SBI Sources of bilingual information 1

QE Quality estimation 1

CAT Compputer aided translation 1

WER Word error rate 2

TER Translation edit rate 2

HTER Human-targeted translation edit rate 2

MT Machine translation 3

IMT Interactive machine translation 4

SMT Statisical machine translation 4

TM Translation memory 5

TU Translation unit 5

FMS Fuzzy-match score 5

TL Target language 6

AER Alignment error rate 160

Bibliography

Albrecht, J. and Hwa, R. (2007). Regression for sentence-level MT evaluation with pseudo references. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 880–887, Stroudsburg, PA, USA.

Allen, J. and Hogan, C. (2000). Toward the development of a post editing module for raw machine translation output: A controlled language perspective. In *3rd International Controlled Language Applications Workshop*, pages 62–71, Washington, DC, USA.

Axelrod, A. (2014). *Data Selection for Statistical Machine Translation*. PhD thesis, University of Washington, Seattle, WA, USA.

Banerjee, P., Rubino, R., Roturier, J., and van Genabith, J. (2015). Quality estimation-guided supplementary data selection for domain adaptation of statistical machine translation. *Machine Translation*, 29(2):77–100.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, volume 29, pages 65–72, Ann Arbor, MI, USA.

Béchara, H., Rubino, R., He, Y., Ma, Y., and van Genabith, J. (2012). An evaluation of statistical post-editing systems applied to RBMT and SMT systems. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 5–230, Mumbai, India.

Biçici, E. (2013). Referential translation machines for quality estimation. In *Proceedings of the 8th Workshop on Statistical Machine Translation*, pages 343–351, Sofia, Bulgaria.

Biçici, E. and Way, A. (2014). Referential translation machines for predicting translation quality. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pages 313–321, Baltimore, MD, USA.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. (2003). Confidence estimation for machine translation. Technical

Report Final Report of the Summer Workshop, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. (2004). Confidence estimation for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, MD, USA.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015). Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the 10th Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal.

Bowker, L. (2002). *Computer-aided translation technology: a practical introduction*, chapter Translation-memory systems, pages 92–127. University of Ottawa Press.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Camargo de Souza, J. G., González-Rubio, J., Buck, C., Turchi, M., and Negri, M. (2014). FBK-UPV-UEdin participation in the WMT14 quality estimation shared-task. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pages 322–328, Baltimore, MD, USA.

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy.

Chander, I. (1998). *Automated Postediting of Documents*. PhD thesis, University of Southern California, Los Angeles, CA, USA.

Chen, J. and Nie, J.-Y. (2000). Automatic construction of parallel English–Chinese corpus for cross-language information retrieval. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, pages 21–28, Seattle, WA, USA.

Désilets, A., Farley, B., Stojanovic, M., and Patenaude, G. (2008). WeBiText: Building large heterogeneous translation memories from parallel web content. In *Proceedings of Translating and the Computer*, pages 27–28, London, UK.

Doddington, G. (2002). Automatic evaluation of machine translation quality using *n*-gram co-occurrence statistics. In *Proceedings of the 2nd International Conference on Human Language Technology Research*, pages 138–145, San Diego, CA, USA.

Duda, R., Hart, P., and Stork, D. (2000). *Pattern Classification*. John Wiley and Sons Inc., second edition.

Eisele, A. and Chen, Y. (2010). MultiUN: A multilingual corpus from united nation documents. In *Proceedings of the 7th Conference on International Language Resources and Evaluation*, pages 2868–2872, Valletta, Malta.

Esplà-Gomis, M. and Forcada, M. L. (2010). Combining content-based and URL-based heuristics to harvest aligned bitexts from multilingual sites with bitextor. *The Prague Bulletin of Mathematical Linguistics*, 93:77–86.

Esplà-Gomis, M., Sánchez-Martínez, F., and Forcada, M. L. (2012). UAlacant: Using online machine translation for cross-lingual textual entailment. In **SEM 2012: The 1st Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main Conference and the Shared Task, and Volume 2: Proceedings of the 6th International Workshop on Semantic Evaluation*, pages 472–476, Montréal, Canada.

Fields, P., Hague, D., Koby, G., Lommel, A., and Melby, A. (2014). What is quality? a management discipline and the translation industry get acquainted. *Tradumàtica*, (12):404–412.

Forcada, M. L. and Sánchez-Martínez, F. (2015). A general framework for minimizing translation effort: towards a principled combination of translation technologies in computer-aided translation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 27–34, Antalya, Turkey.

Foster, G. (2002). *Text prediction for translators*. PhD thesis, Universite de Montreal, Monreal, Canada.

Fung, P. and Cheung, P. (2004). Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM. In *Proceedings of the 2004 Conference on Empirical Method in Natural Language Processing*, pages 57–63, Barcelona, Spain.

Gamon, M., Aue, A., and Smets, M. (2005). Sentence-level MT evaluation without reference translations: Beyond language modeling. In *Proceedings of 9th Annual Meeting of the European Association for Machine Translation*, pages 103–111, Budapest, Hungary.

Gandrabur, S. and Foster, G. (2003). Confidence estimation for translation prediction. In *Proceedings of the 7th Conference on Natural Language Learning*, pages 95–102, Edmonton, Canada.

House, J. (1997). *Translation quality assessment: A model revisited*, volume 410. Gunter Narr Verlag.

Hutchins, J. (2001). Machine translation and human translation: in competition or in complementation? *International Journal of Translation*, 13(1-2):5–20.

Koby, G. S., Melby, A., Fields, P., Lommel, A., and Hague, D. R. (2014). Defining translation quality. *Tradumàtica*, (12):413–420.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand.

Koehn, P. (2009). A web-based interactive computer aided translation tool. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 17–20, Suntec, Singapore.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

Koehn, P., Axelrod, A., Mayne, A. B., Callison-Burch, C., Osborne, M., and Talbot, D. (2005). Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation*, Pittsburgh, USA.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180, Prague, Czech Republic.

Koehn, P., Och, F., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Edmonton, Canada.

Kranias, L. and Samiotou, A. (2004). Automatic translation memory fuzzy match post-editing: a step beyond traditional TM/MT integration. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 331–334, Lisbon, Portugal.

Kuhn, R., Goutte, C., Isabelle, P., and Simard, M. (2011). Method and system for using alignment means in matching translation. USA patent application: US20110093254 A1.

Kuhn, R., Isabelle, P., Goutte, C., Senellart, J., Simard, M., and Ueffing, N. (2010). Automatic post-editing. *Multilingual*, 21(1):43–46.

Lagoudaki, E. (2008). The value of machine translation for the professional translator. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas*, pages 262–269, Waikiki, HI, USA.

Lambert, P., De Gispert, A., Banchs, R., and Mariño, J. B. (2005). Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation*, 39(4):267–285.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Ma, X. and Liberman, M. (1999). Bits: A method for bilingual text search over the Web. In *Proceedings of the 7th Machine Translation Summit*, pages 538–542, Singapore, Singapore.

Melby, A., Fields, P., Koby, G. S., Lommel, A., and Hague, D. R. (2014). Defining the landscape of translation. *Tradumàtica*, (12):392–403.

Mohler, M. and Mihalcea, R. (2008). Babylon parallel text builder: Gathering parallel texts for low-density languages. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 1228–1231, Marrakech, Morocco.

Munteanu, D. and Marcu, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 81–88, Sydney, Australia.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7(4):308–313.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Olive, J. (2005). *DARPA/IPTO Proposer Information Pamphlet*. Global autonomous language exploitation (GALE).

Ortega, J. E., Sánchez-Martínez, F., and Forcada, M. L. (2014). Using any machine translation source for fuzzy-match repair in a computer-aided translation setting. In *Proceedings of the 11th Biennial Conference of the Association for Machine Translation in the Americas, vol. 1: MT Rsearchers*, pages 42–53, Vancouver, Canada.

Papavassiliou, V., Prokopidis, P., and Thurmair, G. (2013). A modular open-source focused crawler for mining monolingual and bilingual corpora from the Web. In *Proceedings of the 6th Workshop on Building and Using Comparable Corpora*, pages 43–51, Sofia, Bulgaria.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, USA.

Plitt, M. and Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics*, 93:7–16.

Quirk, C. (2004). Training a sentence-level machine translation confidence measure. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 825–828, Lisbon, Portugal.

Rehm, G. and Uszkoreit, H. (2013). *META-NET Strategic Research Agenda for Multilingual Europe 2020*. Springer Berlin Heidelberg.

Resnik, P. and Smith, N. A. (2003). The Web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Roukos, S., Graff, D., and Melamed, D. (1995). Hansard French/English. Linguistic Data Consortium. Philadelphia, PA, USA.

Rubino, R., Pirinen, T., Esplà-Gomis, M., Ljubešic, N., Ortiz Rojas, S., Papavassiliou, V., Prokopidis, P., and Toral, A. (2015). Abu-MaTran at WMT 2015 translation task: Morphological segmentation and web crawling. In *Proceedings of the 10th Workshop on Statistical Machine Translation*, pages 184–191, Lisbon, Portugal.

San Vicente, I. and Manterola, I. (2012). PaCo2: A fully automated tool for gathering parallel corpora from the Web. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 1–6, Istanbul, Turkey.

Sánchez-Martínez, F. and Carrasco, R. C. (2011). Document translation retrieval based on statistical machine translation techniques. *Applied Artificial Intelligence*, 25(5):329–340.

Sanchis-Trilles, G., Alabau, V., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., Germann, U., González-Rubio, J., Hill, R., Koehn, P., Leiva, L., Mesa-Lao, B., Ortiz-Martínez, D., Saint-Amand, H., Tsoukala, C., and Vidal, E. (2014). Interactive translation prediction versus conventional post-editing in practice: a study with the CasMaCat workbench. *Machine Translation*, 28(3-4):217–235.

Sikes, R. (2007). Fuzzy matching in theory and practice. *MultiLingual*, 18(6):39–43.

Skadiņš, R., Tiedemann, J., Rozis, R., and Deksne, D. (2014). Billions of parallel words for free: Building and using the EU Bookshop corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th biennial conference of the Association for Machine Translation in the Americas Conference*, pages 223–231, Cambridge, MA, USA.

Somers, H. (2003). *Computers and translation: a translator's guide*, chapter Translation memory systems, pages 31–48. John Benjamins Publishing, Amsterdam, Netherlands.

Specia, L., Raj, D., and Turchi, M. (2010). Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50.

Specia, L., Shah, K., De Souza, J. G., and Cohn, T. (2013). QuEst — a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Conference System Demonstrations)*, pages 79–84, Sofia, Bulgaria.

Sridhar, V. K. R., Barbosa, L., and Bangalore, S. (2011). A scalable approach to building a parallel corpus from the Web. In *Interspeech*, pages 2113–2116, Florence, Italy.

Suchomel, V., Pomikálek, J., et al. (2012). Efficient web crawling for large text corpora. In *Proceedings of the 7th Web as Corpus Workshop*, pages 39–43, Lyon, France.

Tiedemann, J. (2009). News from OPUS — a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Torregrosa, D., Forcada, M. L., and Pérez-Ortiz, J. A. (2014). An open-source web-based tool for resource-agnostic interactive translation prediction. *The Prague Bulletin of Mathematical Linguistics*, 102(1):69–80.

Toselli, A., Vidal, E., and Casacuberta, F. (2011). Interactive machine translation. In *Multimodal Interactive Pattern Recognition and Applications*, pages 135–152. Springer London.

Tyers, F. M. and Alperen, M. S. (2010). South-east european times: A parallel corpus of Balkan languages. In *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, pages 49–53, Valletta, Malta.

Ueffing, N. and Ney, H. (2005). Application of word-level confidence measures in interactive statistical machine translation. In *Proceedings of the 10th European Association for Machine Translation Conference "Practical applications of machine translation"*, pages 262–270, Budapest, Bulgaria.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing*, pages 590–596, Borovets, Bulgaria.

Zhao, B. and Vogel, S. (2002). Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, pages 745–748, Washington, DC, USA.