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

Miquel Esplà-Gomis

Universitat d'Alacant
Departament de Llenguatges i Sistemes Informàtics

January 25, 2016

About the title

Quite long title for a dissertation; let's dissect it:

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

Translation technologies

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

Translation technologies

Technologies that provide translators with translation proposals (hypotheses) for a SL segment to ease the translation task:

- machine translation (MT)
- translation memories (TM)
- interactive machine translation
- automatic post-editing
- etc.

Translation technologies

Technologies that provide translators with translation proposals (hypotheses) for a SL segment to ease the translation task:

- **machine translation (MT)**
- **translation memories (TM)**
- interactive machine translation
- automatic post-editing
- etc.

Word-level quality estimation

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

Word-level quality estimation

Quality estimation: predicting the quality of a translation hypothesis without using preexisting reference translations

- helpful to estimate translation effort and, therefore, for budgeting
- allows choosing the most helpful translation technology

Word-level quality estimation: predicting the quality of each word in a translation hypothesis

- it can guide the translator during post-editing

External sources of bilingual information

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

External sources of bilingual information

Sources of bilingual information (SBI): any source of information that allows to translate sub-segments:

- machine translation
- translation memories
- phrase tables
- etc.

External: sources of information that are independent of the translation technologies to be evaluated and can be used as black boxes

Where are we?

Where are we?

Computer aided-translation (CAT) environments provide **different translation technologies** to help translators...

Where are we?

Computer aided-translation (CAT) environments provide **different translation technologies** to help translators...

... some CAT environments provide **word-level QE** but it is only for **MT**...

Where are we?

Computer aided-translation (CAT) environments provide **different translation technologies** to help translators...

... some CAT environments provide **word-level QE** but it is only for **MT**...

... there are many approaches for word-level QE, but they are based on **specifically built resources** which usually require to be **built independently**.

Where would we like to be?

Where would we like to be?

We would like CAT environments in which **word-level QE** is provided **for any translation technology** available...

Where would we like to be?

We would like CAT environments in which **word-level QE** is provided **for any translation technology** available...

... **without** having to build any **specific resources**...

Where would we like to be?

We would like CAT environments in which **word-level QE** is provided **for any translation technology** available...

... **without** having to build any **specific resources**...

... that only require to plug **any SBI already available**, either locally or on the Internet...

Where would we like to be?

We would like CAT environments in which **word-level QE** is provided **for any translation technology** available...

... **without** having to build any **specific resources**...

... that only require to plug **any SBI already available**, either locally or on the Internet...

... and that are able to **build their own SBI** if needed.

Research steps

1. Definition of methods for word-level QE for TM-based CAT using SBI
   - new problem identified
   - more work devoted
   - good starting point: more information available than in MT

Research steps

1. Definition of methods for word-level QE for TM-based CAT using SBI
   - new problem identified
   - more work devoted
   - good starting point: more information available than in MT
2. Extending these methods to MT
   - important differences between both technologies
   - the two most important translation technologies are covered

Research steps

1. Definition of methods for word-level QE for TM-based CAT using SBI
   - new problem identified
   - more work devoted
   - good starting point: more information available than in MT
2. Extending these methods to MT
   - important differences between both technologies
   - the two most important translation technologies are covered
3. Researching on methods to create SBI for under-resourced language pairs
   - developed in parallel
   - improves the usability of the word-level QE methods developed

Main working hypothesis

It is possible to develop methods exclusively **based on external SBI** for **word-level QE** in **TM-based (CAT) and MT**.

Outline

Outline

Outline

Translation memories

Catalan	English
S_1: L' EAMT és membre de l' IAMT	T_1: The EAMT is a member of the IAMT
S_2: L' Associació Internacional per a la Traducció Automàtica	T_2: The International Association for Machine Translation
S_3: el congrés d' enguany se celebra a Lovaina	T_3: current year 's conference is held in Leuven
...	...

Translation memories

Catalan	English
S_1: L' EAMT és membre de l' IAMT	T_1: The EAMT is a member of the IAMT
S_2: L' Associació Internacional per a la Traducció Automàtica	T_2: The International Association for Machine Translation
S_3: el congrés d' enguany se celebra a Lovaina	T_3: current year 's conference is held in Leuven
...	...

S': Associació Europea per a la Traducció Automàtica

Translation memories

Catalan	English
S_1: L' EAMT és membre de l' IAMT	T_1: The EAMT is a member of the IAMT
S_2: L' Associació Internacional per a la Traducció Automàtica	T_2: The International Association for Machine Translation
S_3: el congrés d' enguany se celebra a Lovaina	T_3: current year 's conference is held in Leuven
...	...

S': Associació Europea per a la Traducció Automàtica

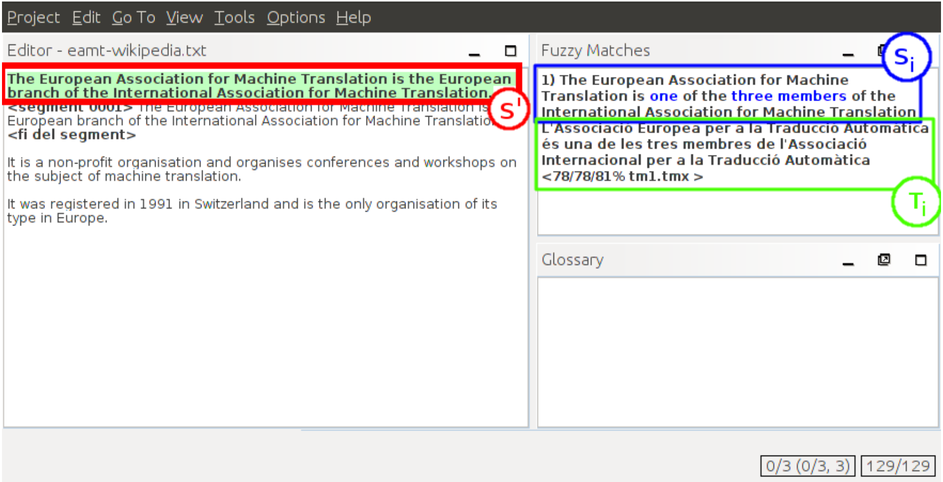
S_2: L' Associació Internacional per a la Traducció Automàtica

Translation memories

Catalan	English
S_1: L' EAMT és membre de l' IAMT	T_1: The EAMT is a member of the IAMT
S_2: L' Associació Internacional per a la Traducció Automàtica	T_2: The International Association for Machine Translation
S_3: el congrés d' enguany se celebra a Lovaina	T_3: current year 's conference is held in Leuven
…	…

S': Associació Europea per a la Traducció Automàtica
S_2: L' Associació Internacional per a la Traducció Automàtica
T_2: The International Association for Machine Translation

Translation-memory-based CAT tools



Word-level QE for TM-based CAT

The problem: detecting which words in suggestions provided by TM-based CAT tools should be post-edited and which should be kept

Example

S' = "Associació Europea per a la Traducció Automàtica"
S_2 = "L' Associació Internacional per a la Traducció Automàtica"
T_2 = "The International Association for Machine Translation"
T'="European Association for Machine Translation"

Outline

Word-level QE for MT

The problem: detecting which words in MT output should be post-edited

Example

S'="Associació Europea per a la Traducció Automàtica"

MT(S')="European Association for the Automatic Translation"

T'="European Association for Machine Translation"

Word-level QE for MT

The problem: detecting which words in MT output should be post-edited

Example

S'=“Associació Europea per a la Traducció Automàtica”

MT(S')=“European Association for the Automatic Translation”

T'=“European Association for Machine Translation”

Word-level QE: MT vs. TM-based CAT

T provided by **MT** for S': T is a translation of S'
that may be adequate or not

T provided by **TM-based CAT** for S': T is not a translation of S',
but is an adequate translation of a similar segment S

Outline

1. Introduction to the problems addressed
   - Word-level QE in TM-based CAT
   - Word-level QE in MT
   - Building new SBI for under-resourced language pairs
2. Word-level QE in TM-based CAT
3. Word-level QE in MT
4. Building SBI for under-resourced language pairs
5. Contributions to the state-of-the-art

Creating new SBI for under-resourced languages

The scenario: no SBI available for a translation task between under-resourced languages

The solution: building new SBI by using the Web as a parallel corpus

- A parallel data crawler can be used to build a parallel corpus
- Many SBI can be automatically built from a parallel corpus:
  - dictionaries
  - MT systems
  - bilingual concordancers
  - etc.

The Web as a parallel corpus

Outline

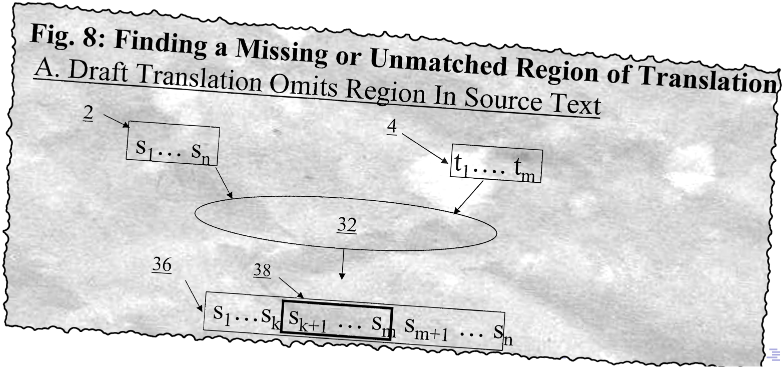
Related work

- **No previous works** on word-level QE in TM-based CAT...

Related work

- **No previous works** on word-level QE in TM-based CAT...

- ...but the patent application by **Kuhn et al. (2011)**: word-level QE in TM-based CAT by using statistical word alignments **(no details)**



Related work

- **No previous works** on word-level QE in TM-based CAT...
- ...but the patent application by **Kuhn et al. (2011)**: word-level QE in TM-based CAT by using statistical word alignments **(no details)**
- Closest approach by **Kranias and Samioutou (2004)**: automatic post-editing of translation suggestions using glossaries for word-alignment and MT

Word-level QE in TM-based CAT: a new problem

- **word-level QE** is a concept that **originates in MT**

Word-level QE in TM-based CAT: a new problem

- **word-level QE** is a concept that **originates in MT**
- one of the keys of success of TM-based CAT tools:
  segment-level QE with easily-interpretable fuzzy-match scores

Word-level QE in TM-based CAT: a new problem

- **word-level QE** is a concept that **originates in MT**
- one of the keys of success of TM-based CAT tools: **segment-level QE** with easily-interpretable fuzzy-match scores
- **word-level QE** for TM-based CAT is a new problem: is it relevant?

Word-level QE in TM-based CAT: a new problem

- **word-level QE** is a concept that **originates in MT**
- one of the keys of success of TM-based CAT tools: **segment-level QE** with easily-interpretable fuzzy-match scores
- **word-level QE** for TM-based CAT is a new problem: is it relevant?

YES: pilot study indicates that word-level QE for TM-based CAT could improve productivity of translators up to 14%

Rationale

Edit distance provides information about the words in S that do not appear in S':

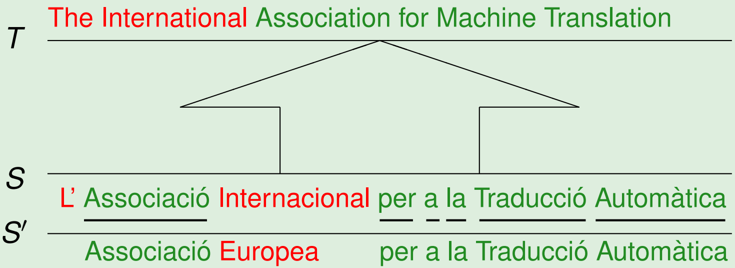
Example

T	The International Association for Machine Translation
S	L' Associació Internacional per a la Traducció Automàtica
S'	Associació Europea per a la Traducció Automàtica

Rationale

The objective is to project source-side matching information onto T, to suggest which words to change (bad) or keep unedited (good):

Example



Outline

Working hypothesis #1

H1: It is possible to use **word alignments** to **estimate the quality** of **TM-based CAT suggestions** at the word level

Rationale

[BAD] [GOOD]

T The International Association for Machine Translation

S L' Associació Internacional per a la Traducció Automàtica

S′ Associació Europea per a la Traducció Automàtica

- *Automàtica* and *Machine* **aligned** & *Automàtica* **matched** $\Longrightarrow$ **GOOD**
- *Internacional* and *International* **aligned** & *Internacional* **unmatched** $\Longrightarrow$ **BAD**

Rationale

[?] [?]

T The International Association for Machine Translation

?

S L' Associació Internacional per a la Traducció Automàtica

S' Associació Europea per a la Traducció Automàtica

- *for* **not aligned** $\Longrightarrow$ **???**
- *The* **contradictory evidence** $\Longrightarrow$ **???** (voting scheme)

Outline

1. Introduction to the problems addressed
2. Word-level QE in TM-based CAT
   - Word-level QE in TM-based CAT using statistical word alignments
   - Word-level QE in CAT using SBI-based word alignments
   - Word-level QE in TM-based CAT directly using SBI
   - Evaluation
3. Word-level QE in MT
4. Building SBI for under-resourced language pairs
5. Contributions to the state-of-the-art

Working hypothesis #2

H2: It is possible to use any **SBI to obtain word alignments**

Using SBI to obtain word-alignments

Using sub-segment alignments for word alignments based on SBI

Example

S: L' Associació Internacional per a la Traducció Automàtica (Catalan)
T: The International Association for Machine Translation (English)

Segmentation

S: L' Associació Internacional per a la Traducció Automàtica (Catalan)
T: The International Association for Machine Translation (English)

Catalan

L'
Associació
Internacional
per
a
la
Traducció
Automàtica
L'Associació
Associació Internacional
Internacional per
...

English

The
International
Association
for
Machine
Translation
The International
International Association
Association for
for Machine
...

Sub-segment translation to obtain pairs (σ, τ)

S: L' Associació Internacional per a la Traducció Automàtica (Catalan)
T: The International Association for Machine Translation (English)

Catalan → English

L' → The
Associació → Association
Internacional → International
per → by
a → at
la → the
Traducció → Translation
Automàtica → Automatic
L'Associació → The Association
Associació Internacional → International Association
Internacional per → International for
...

English → Catalan

The → El
International → Internacional
Association → Associació
for → per a
Machine → Màquina
Translation → Traducció
The International → El Internacional
International Association → Associació Internacional
Association for → Associació per
for Machine → per a Màquina
...

(σ, τ) filtering

S: L' Associació Internacional per a la Traducció Automàtica (Catalan)
T: The International Association for Machine Translation (English)

Catalan → English

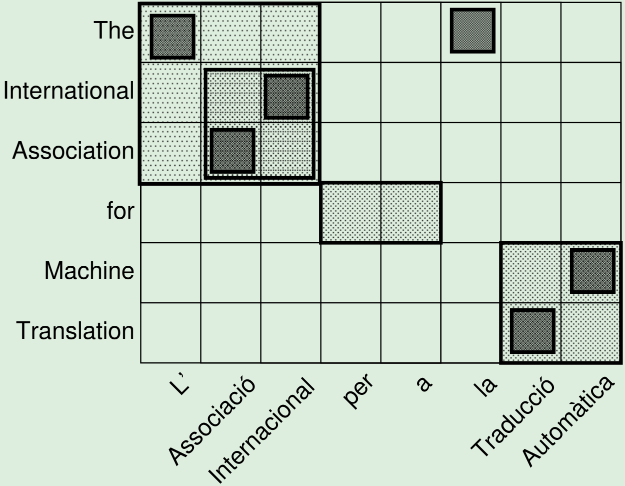
L' → The
Associació → Association
Internacional → International
~~per → by~~
~~a → at~~
la → the
Traducció → Translation
~~Automàtica → Automatic~~
~~L'Associació → The Association~~
Associació Internacional → International Association
~~Internacional per → International for~~
...

English → Catalan

~~The → El~~
International → Internacional
Association → Associació
for → per a
~~Machine → Màquina~~
Translation → Traducció
~~The International → El Internacional~~
International Association → Associació Internacional
~~Association for → Associació per~~
~~for Machine → per a Màquina~~
...

Sub-segment alignment

Example



Using SBI to obtain word-alignments?

1. **Segmentation** of both sentences
2. **Translation** of sub-segments (in both directions) using SBIs to obtain sub-segment pairs (σ, τ)
3. **Filtering** pairs (σ, τ) with σ not in S or τ not in T
4. **Aligning sub-segments** (σ, τ) between S and T

Using SBI to obtain word-alignments?

1. **Segmentation** of both sentences
2. **Translation** of sub-segments (in both directions) using SBIs to obtain sub-segment pairs (σ, τ)
3. **Filtering** pairs (σ, τ) with σ not in S or τ not in T
4. **Aligning sub-segments** (σ, τ) between S and T
5. **Aligning words** by using sub-segment alignments

From sub-segment alignments to word alignments

Option A: Intuitive heuristic method based on the **concept of pressure**:

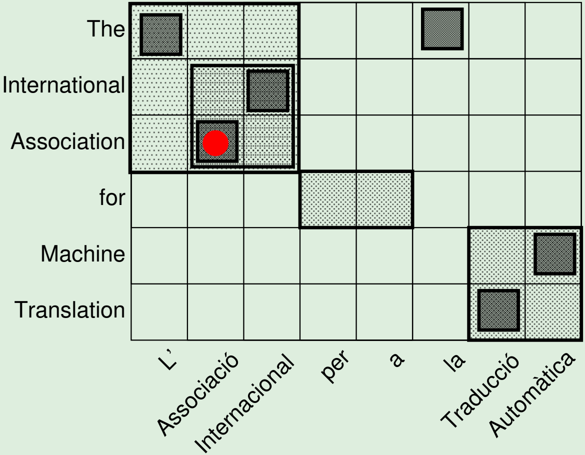
From sub-segment alignments to word alignments

Option A: Intuitive heuristic method based on the **concept of pressure**:

- Every sub-segment alignment **applies force of 1**;
- Every sub-segment alignment applies, on each pair of words, a **pressure** corresponding to its **force divided by its area**;
- **Alignment strength between two words**: summation of **all the pressures on a pair of words.**

Word alignment strength

Example



Word-alignment strength between ***Associació*** and ***Association***:

$$\frac{1}{1\times1} + \frac{1}{2\times2} + \frac{1}{3\times3} \simeq 1.36$$

Word alignment strength

Example

	L’	Associació	Internacional	per	a	la	Traducció	Automàtica
The	1.11	0.11	0.11			1		
International	0.11	0.36	1.36					
Association	0.11	1.36	0.36					
for				0.5	0.5			
Machine							0.25	1.25
Translation							1.25	0.25

Word alignment strength

Example

	L'	Associació	Internacional	per	a	la	Traducció	Automàtica
The	1.11	0.11	0.11			1		
International	0.11	0.36	1.36					
Association	0.11	1.36	0.36					
for				0.5	0.5			
Machine							0.25	1.25
Translation							1.25	0.25

From sub-segment alignments to word alignments

Option B: Linear-combination approach:

From sub-segment alignments to word alignments

Option B: Linear-combination approach:

Sub-segment alignments are grouped **depending on their geometry** (1×1, 1×2, 2×1, etc.), and a **linear-combination** approach is used to **assign weights** to them

Outline

1. Introduction to the problems addressed

2. Word-level QE in TM-based CAT
   - Word-level QE in TM-based CAT using statistical word alignments
   - Word-level QE in CAT using SBI-based word alignments
   - Word-level QE in TM-based CAT directly using SBI
   - Evaluation

3. Word-level QE in MT

4. Building SBI for under-resourced language pairs

5. Contributions to the state-of-the-art

Working hypothesis #3

H3: It is possible to use **SBI to estimate the quality** of TM-based CAT translation suggestions at the word level

Rationale

- Align sub-segments between *S* and *T* (as in the method using SBI-based word alignment)
- Obtain word-level QEs by:
  - using an heuristic method that counts matched/unmatched aligned words (as in the method using word alignments)
  - using a binary classifier to make quality estimations

Evidence from sub-segments

Evidence from sub-segments of length $L = 1$

Example

T The International Association for Machine Translation

S L' Associació Internacional per a la Traducció Automàtica

S' Associació Europea per a la Traducció Automàtica

Contradictory evidence from sub-segments

Contradictory evidence from sub-segments of length $L = 1$

Example

T The International Association for Machine Translation

S L' Associació Internacional per a la Traducció Automàtica

S' Associació Europea per a la Traducció Automàtica

Evidence from longer sub-segments

Longer sub-segments provide more context

Example

T The International Association for Machine Translation

S L' Associació Internacional per a la Traducció Automàtica

S' Associació Europea per a la Traducció Automàtica

Outline

1. Introduction to the problems addressed

2. Word-level QE in TM-based CAT
   - Word-level QE in TM-based CAT using statistical word alignments
   - Word-level QE in CAT using SBI-based word alignments
   - Word-level QE in TM-based CAT directly using SBI
   - Evaluation

3. Word-level QE in MT

4. Building SBI for under-resourced language pairs

5. Contributions to the state-of-the-art

Evaluation of all the approaches

Evaluation for the **five approaches**:

- statistical word alignment
- SBI-based word alignment ("pressure")
- SBI-based word alignment (linear combination)
- SBI directly (heuristic)
- SBI directly (binary classification)

Metrics: accuracy (A) and fraction of **words not covered** (NC)

Evaluation resources

- TM and test set from the **same domain** for **Spanish→English** (ES→EN)
- Different values of **fuzzy-match score** threshold (Θ)
- **Three SBI (MT)**: Google Translate, Apertium, and Power Translator

Using SBI-based word-alignments

training: ES→EN; in-domain; all SBI **test:** ES→EN; all SBI

method	metric	$\Theta \geq 60\%$	$\Theta \geq 70\%$	$\Theta \geq 80\%$	$\Theta \geq 90\%$
statistical word alignment	$A(\%)$	93.9±.2	94.3±.2	95.1±.2	95.3±.3
	$\mathrm{NC}(\%)$	4.4±.1	4.5±.2	4.3±.2	4.4±.3

Using SBI-based word-alignments

training: ES→EN; in-domain; all SBI **test:** ES→EN; all SBI

method	metric	$\Theta \geq 60\%$	$\Theta \geq 70\%$	$\Theta \geq 80\%$	$\Theta \geq 90\%$
statistical word alignment	$A(\%)$	93.9±.2	94.3±.2	95.1±.2	95.3±.3
	$\mathrm{NC}(\%)$	**4.4±.1**	**4.5±.2**	**4.3±.2**	**4.4±.3**
pressure SBI-based word alignment	$A(\%)$	93.7±.2	94.5±.2	95.4±.2	96.0±.2
	$\mathrm{NC}(\%)$	10.0±.2	8.9±.2	8.2±.3	7.2±.3
linear SBI-based word alignment	$A(\%)$	94.2±.2	94.8±.2	95.8±.2	96.4±.2
	$\mathrm{NC}(\%)$	9.8±.2	8.8±.2	8.1±.3	7.0±.3

Using SBI-based word-alignments

training: ES→EN; in-domain; all SBI **test:** ES→EN; all SBI

method	metric	$\Theta \geq 60\%$	$\Theta \geq 70\%$	$\Theta \geq 80\%$	$\Theta \geq 90\%$
statistical word alignment	$A(\%)$	93.9±.2	94.3±.2	95.1±.2	95.3±.3
	$\mathrm{NC}(\%)$	**4.4±.1**	**4.5±.2**	**4.3±.2**	**4.4±.3**
pressure SBI-based word alignment	$A(\%)$	93.7±.2	94.5±.2	95.4±.2	96.0±.2
	$\mathrm{NC}(\%)$	10.0±.2	8.9±.2	8.2±.3	7.2±.3
linear SBI-based word alignment	$A(\%)$	94.2±.2	94.8±.2	95.8±.2	96.4±.2
	$\mathrm{NC}(\%)$	9.8±.2	8.8±.2	8.1±.3	7.0±.3
heuristic method using SBI directly	$A(\%)$	93.3±.2	93.9±.2	95.1±.2	96.5±.3
	$\mathrm{NC}(\%)$	5.1±.1	5.2±.2	5.5±.2	5.9±.3
binary classification using SBI directly	$A(\%)$	**95.1±.1**	**95.6±.2**	**96.4±.2**	**96.9±.2**
	$\mathrm{NC}(\%)$	5.1±.1	5.2±.2	5.5±.2	5.9±.3

QE accuracy across domains

- Training on **out-of-domain TMs** and re-using the models
- **Important**: simulates performance on a translation proposal (translation unit) not seen during training
- Comparing the method based on **statistical word alignment** and methods using **SBI directly**

QE accuracy across domains

training: ES→EN; out-of-domain; all SBI **test:** ES→EN; all SBI

Θ(%)	training domain	word alignment statistical A (%)	word alignment statistical NC (%)	SBI-based approaches classifier A (%)	SBI-based approaches heuristic A (%)	SBI-based approaches NC (%)
≥ 60	in domain	93.9±.2	6.1±.2	95.1±.1		
	out of domain 1	91.8±.2	9.6±.2	**93.3±.2**	**93.3±.2**	**5.1±.1**
	out of domain 2	90.2±.2	11.7±.2	92.0±.2		
≥ 70	in domain	94.3±.2	5.9±.2	95.6±.2		
	out of domain 1	92.7±.2	9.7±.2	**94.0±.2**	**94.1±.2**	**5.2±.2**
	out of domain 2	91.6±.2	11.6±.3	92.7±.2		
≥ 80	in domain	95.1±.2	5.4±.2	96.4±.2		
	out of domain 1	93.8±.2	9.2±.3	**95.6±.2**	95.3±.2	**5.5±.2**
	out of domain 2	93.2±.2	11.1±.3	94.3±.2		
≥ 90	in domain	95.3±.3	4.9±.3	96.9±.2		
	out of domain 1	94.6±.3	8.9±.3	**96.5±.2**	**96.6±.2**	**5.9±.3**
	out of domain 2	94.2±.3	10.3±.4	96.3±.2		

QE accuracy across SBI

- Training on a translation task using **Apertium** and **Power Translator** separately as SBI
- Evaluating on the same translation task using **Google Translate** as SBI
- Comparing the methods using **SBI directly**

QE accuracy across SBI

training: ES→EN; in-domain; SBI ≠ Google **test:** ES→EN; SBI = Google

$\Theta(\%)$	A (%) binary classification using SBI directly			A (%) heuristic method
	Apertium	**Google Translate**	**Power Translator**	**using SBI directly**
≥ 60	92.9±.2	95.1±.1	91.1±.2	**93.4**±**.2**
≥ 70	93.4±.2	95.5±.2	93.0±.2	**94.3**±**.2**
≥ 80	**95.6**±**.2**	96.3±.2	94.8±.2	95.4±.2
≥ 90	96.7±.2	97.0±.2	96.5±.2	96.7±.2

QE accuracy across language pairs

- Impact of re-using models for different languages
- Models trained on **9 different language pairs**: EN→ES, DE→EN, EN→DE, FR→EN, EN→FR, FI→EN, EN→FI, ES→FR, FR→ES
- Evaluation on **ES→EN**
- Comparing the methods using **SBI directly**

QE accuracy across language pairs: **impact of the TL**

training: XX→YY; in-domain; all SBI **test:** ES→EN; all SBI

to EN	$\Theta = 60\%$	**from EN**	$\Theta = 60\%$
ES→EN	$95.1{\pm}.1$	EN→ES	$90.8{\pm}.2$
DE→EN	$92.5{\pm}.2$	EN→DE	$92.2{\pm}.2$
FR→EN	$92.7{\pm}.2$	EN→FR	$90.9{\pm}.2$
FI→EN	$92.4{\pm}.2$	EN→FI	$91.1{\pm}.2$

QE accuracy across language pairs: **best results**

training: XX→YY; in-domain; all SBI **test:** ES→EN; all SBI

training language pair	$\Theta(\%)$ ≥ 60	≥ 70	≥ 80	≥ 90
ES→EN	95.1±.1	95.5±.2	96.3±.2	97.1±.2
FR→EN	92.7±.2	**94.2±.2**	**95.8±.2**	**96.7±.2**
heuristic	**93.4±.2**	**94.3±.2**	95.4±.2	**96.7±.2**

Chapter conclusions

- **New problem** identified that is relevant for professional translators
- New methods for **word alignments based on SBI**
- Three methods for word level QE for CAT
  - **Statistical word alignment**: best coverage (in-domain), difficult to re-use models when domain changes
  - **SBI-based word alignment**: worst coverage
  - **SBI directly for word-level QE**: best accuracy, good results when re-using models across domains

Outline

Related work

- Confidence estimation (Blatz et al., 2003): semantic features (WordNet), probabilistic lexicons, word posterior probabilities
- Approaches using **pseudo-references**:
  - word occurrence in *n*-best lists (Ueffing and Ney, 2007)
  - pseudo-references (Biçici, 2013)
  - inverse machine translation (MULTILIZER approach: Bojar et al., 2014)
- General framework for translation QE in MT: **QuEst** (Specia, 2013) [now **QuEst++** (Specia, 2015)]

Working hypothesis #4

H4: 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**

Adapting the existing methods?

- We start with the **binary classification approach** based on SBI
- We can reuse the idea of **sub-segment alignments**...
- ... but we do not have S and T: it is not possible to use **exactly the same method**

Adapting the existing methods?

New collection of features is needed

- **positive** features using **exact matches** of translated sub-segments
- **negative** features using **partial matches** of translated sub-segments

Example

S'="Associació Europea per a la Traducció Automàtica"

MT(S')="European Association for the Automatic Translation"

T'="European Association for Machine Translation"

How to use SBI to obtain positive evidence?

1. Source language sub-segments σ and target language sub-segments τ are translated using SBI
2. Words in a sub-segment τ that is related by SBI to a sub-segment σ that matches S' get positive evidence

S' = "Associació Europea per a la Traducció Automàtica"

translate σ

τ = "European Association for"

✓ ✓ ✓

$\mathrm{MT}(S')$ = "European Association for the Automatic Translation"

How to use SBI to obtain negative evidence?

1. Source language sub-segments σ are translated using SBI
2. Translations τ are partially matched onto $\mathrm{MT}(S')$
3. Words in $\mathrm{MT}(S')$ that do not match but are surrounded by matching words get negative evidence

Example

S' = "Associació Europea per a la Traducció Automàtica"

translate σ

τ = the Machine Translation

✓ ✗ ✓

$\mathrm{MT}(S')$ = "European Association the Automatic Translation"

Outline

Experimental setting

- Binary classification problem: **GOOD** and **BAD** classes
- Metrics: F_1^w, F_1^{BAD}, and F_1^{GOOD}
- Datasets:
  - **WMT'15**: we **did** compete; 1 dataset (for translating Spanish into English)
  - **WMT'14**: we **did not** compete; 4 datasets (for 4 language pairs)

SBI used

- *Machine translation*: **Google Translate** and **Apertium**
  - only one translation suggestion for each σ
  - translation frequency not available

- *Bilingual concordancer*: **Reverso Context**
  - multiple translation suggestions possible for each σ
  - translation frequency available

WMT'15 participant ranking

WMT'15 evaluation on EN→ES data

System ID	F_1^w	$\boldsymbol{F_1^{\mathrm{BAD}}}$	F_1^{GOOD}
• UAlacant/OnLine-SBI-Baseline	71.5%	43.1%	78.1%
• HDCL/QUETCHPLUS	72.6%	43.1%	79.4%
UAlacant/OnLine-SBI	69.5%	41.5%	76.1%
SAU/KERC-CRF	77.4%	39.1%	86.4%
SAU/KERC-SLG-CRF	77.4%	38.9%	86.4%
SHEF2/W2V-BI-2000	65.4%	38.4%	71.6%
SHEF2/W2V-BI-2000-SIM	65.3%	38.4%	71.5%
SHEF1/QuEst++-AROW	62.1%	38.4%	67.6%
...			

Results WMT'14

SBI-based methods vs. best performing systems in WMT'14

language pair	OnLine-SBI F_1^w	OnLine-SBI F_1^{BAD}	winner method	F_1^w	F_1^{BAD}
EN→ES	61.4%	49.0%	FBK-UPV-UEDIN/RNN	62.0%	48.7%
ES→EN	75.9%	10.4%	RTM-DCU/RTM-GLMd	79.5%	29.1%
EN→DE	66.8%	43.1%	RTM-DCU/RTM-GLM	71.5%	45.3%
DE→EN	75.0%	40.3%	RTM-DCU/RTM-GLM	72.4%	26.1%

- We are good when detecting errors: *terminology*, *mistranslation*, and *unintelligible*

Chapter conclusions

- **Confirmed**: SBI can be used for word-level QE in MT
- **Leading performer** among the approaches proposed
- **Small amount of features** required (**70**) when compared to other approaches: for example Biçici (2013) uses 511K features
- **Independence with respect to the MT system** used to produce the translation hypotheses

Outline

Rationale

1. Using **Bitextor** (Esplà-Gomis, 2010) to build a parallel corpus for under-resourced language pairs
2. Using the parallel corpus obtained to build new SBI
3. Using the new SBI for the QE techniques developed

Working hypotheses #5 and #6

H5: 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

H6: The **results** obtained in word-level QE for under-resourced language pairs can be **improved by using new SBI** obtained through parallel data crawling

Outline

1. Introduction to the problems addressed
2. Word-level QE in TM-based CAT
3. Word-level QE in MT
4. Building SBI for under-resourced language pairs
   - Using the Web as a parallel corpus
   - Building SBI for English–Croatian
   - New SBI in word-level QE in TM-based CAT
5. Contributions to the state-of-the-art

Related work

Approaches to detect parallel text in multilingual websites:

- **URL similarity patterns**: Ma and Liberman (1999), Nie et al. (1999), and Resnik and Smith (2003)
- **HTML structure comparison**: Resnik and Smith (2003), Sin et al. (2005), and Zhang et al. (2006)
- **image coocurrence**: Papavassiliou et al. (2013)
- **text comparison** (mostly based on bag-of-words overlapping metrics): Chen et al. (2004), and Barbosa et al. (2012)

How does Bitextor work?

HTML structure comparison

```
<div>
    <h1>
        Aquest és un títol
    </h1>
        Aquest és un fragment de text.
    <a href="http://www.anurl.cat/cat">
        <img src="an_image.png">
    </a>
</div>
```

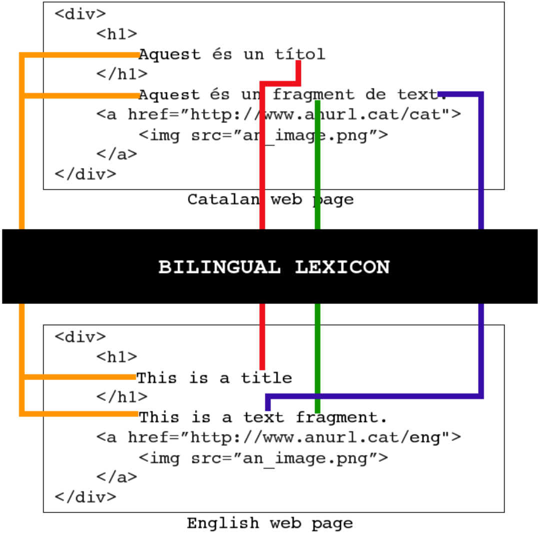
Catalan web page

```
<div>
    <h1>
        This is a title
    </h1>
        This is a text fragment.
    <a href="http://www.anurl.cat/eng">
        <img src="an_image.png">
    </a>
</div>
```

English web page

How does Bitextor work?

Word overlap metrics based on Sánchez-Martínez & Carrasco (2011)



Outline

1. Introduction to the problems addressed
2. Word-level QE in TM-based CAT
3. Word-level QE in MT
4. Building SBI for under-resourced language pairs
   - Using the Web as a parallel corpus
   - Building SBI for English–Croatian
   - New SBI in word-level QE in TM-based CAT
5. Contributions to the state-of-the-art

Building a TM for English–Croatian tourism domain

- Crawling **21 bilingual websites** from the tourism domain
- Using two crawlers: **Bitextor** and **ILSP Focused Crawler**
- For both crawlers, 2 configurations: **accuracy-oriented** and **recall-oriented**
- Evaluation: resulting parallel corpus through **manual revision**

Building a TM for English–Croatian tourism domain

Evaluation of corpora built by Bitextor and ILSP Focused Crawler

tool	setting	aligned documents	success rate	unique aligned segments
Focused Crawler	recall	3,294	73.86%	40,431
	accuracy	2,406	90.76%	32,544
Bitextor	recall	**7,787**	74.70%	**50,338**
	accuracy	3,758	**94.79%**	36,834

Outline

1. Introduction to the problems addressed
2. Word-level QE in TM-based CAT
3. Word-level QE in MT
4. Building SBI for under-resourced language pairs
   - Using the Web as a parallel corpus
   - Building SBI for English–Croatian
   - New SBI in word-level QE in TM-based CAT
5. Contributions to the state-of-the-art

Experimental setup

- Evaluation on **word-level QE in TM-based CAT** for **English–Finnish**
- Parallel corpus crawled using Bitextor on about **10,700 websites** (**1,700,000 parallel segments**)
- Performance when using **two kinds of SBI**: **phrase tables** and **phrase-based SMT** (Moses)
- Comparison to **Google Translate** as a SBI

Results for Finnish→English

Comparing Google with new SBI created with Bitextor

SBI	metric	method	$\Theta \geq 60\%$	$\Theta \geq 70\%$	$\Theta \geq 80\%$	$\Theta \geq 90\%$
Google	$A(\%)$	**classification**	93.2±.2	94.6±.2	94.7±.2	94.8±.3
		heuristic	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	$A(\%)$	**classification**	87.7±.3	90.3±.3	91.8±.3	92.3±.4
		heuristic	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	$A(\%)$	**classification**	93.1±.2	94.7±.2	94.9±.3	94.7±.4
		heuristic	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 + Moses + phrases	$A(\%)$	**classification**	91.3±.2	92.8±.2	93.2±.3	93.0±.3
		heuristic	87.8±.2	90.5±.2	91.8±.3	92.6±.4
	NC(%)	—	**5.1±.2**	**4.3±.2**	**4.9±.2**	**5.4±.3**

Results for Finnish→English

Comparing Google with new SBI created with Bitextor

SBI	metric	method	$\Theta \geq 60\%$	$\Theta \geq 70\%$	$\Theta \geq 80\%$	$\Theta \geq 90\%$
Google	$A(\%)$	**classification**	93.2±.2	94.6±.2	94.7±.2	94.8±.3
		heuristic	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	$A(\%)$	**classification**	87.7±.3	90.3±.3	91.8±.3	92.3±.4
		heuristic	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	$A(\%)$	**classification**	93.1±.2	94.7±.2	94.9±.3	94.7±.4
		heuristic	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 + Moses	$A(\%)$	**classification**	91.3±.2	92.8±.2	93.2±.3	93.0±.3
		heuristic	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**

Chapter conclusions

- Bitextor provides **high-quality parallel corpora**
- Bitextor may **enable the use of SBI-based approaches** for word-level QE in TM-based CAT for under-resourced language pairs
- Encouraging results:
  - Phrase-based SMT system: accuracy comparable to that of Google Translate
  - Phrase tables + Google Translate: coverage improves more than twice; accuracy falls noticeably

Outline

Where would we like to be?

We would like CAT environments in which **word-level QE** is provided **for any translation technology** available...

... **without** having to build any **specific resources**...

... that only require to plug **any SBI already available** for the user...

... and that are able to **build their own SBI** if needed.

How does our work contribute?

How does our work contribute?

We have developed methods for the **2 most used** technologies in CAT, **TM** and **MT**,...

How does our work contribute?

We have developed methods for the **2 most used** technologies in CAT, **TM** and **MT**,...

... that only need to have access to the **SBI already available** for the user...

How does our work contribute?

We have developed methods for the **2 most used** technologies in CAT, **TM** and **MT**,...

... that only need to have access to the **SBI already available** for the user...

... and if **no SBI are available**, we have developed a methodology for **creating new SBI from multilingual websites** using the tool **Bitextor**.

New research paths

- **QE for parallel-corpora crawling**
- **Automatic or aided post-editing**
- **Word-level QE for interactive machine translation**

QE for parallel-corpora crawling

- **QE for parallel-corpora crawling**
  - QE as a new source of information for comparing documents
  - Using word-level QE as a source of information for cleaning parallel data
- Automatic or aided post-editing
- Word-level QE for interactive machine translation

Automatic or aided post-editing

- QE for parallel-corpora crawling
- **Automatic or aided post-editing**
    - Adapting current techniques to suggest translation alternatives
    - Extend the current work to consider insertions
- Word-level QE for interactive machine translation

Word-level QE for interactive machine translation

- QE for parallel-corpora crawling
- Automatic or aided post-editing
- **Word-level QE for interactive machine translation**
  - Using the SBI-based word-alignment methods developed to keep track of the sub-segments translated
  - Evaluating word-level QE for this new technology

Free/open-source software published

Software published (`http://bit.do/mespla_software`):

- `Gamblr-CAT`: Software for word-level QE in TM-based CAT
- `Gamblr-MT`: Software for word-level QE in MT
- `Bitextor`: Software for building parallel data from multilingual websites
- `OmegaT-SessionLog`: Plugin for traking the actions of a translator when using the TM-based CAT tool OmegaT
- `OmegaT-Marker-Plugin`: Plugin for OmegaT that implements the heuristic word-level QE system
- `Flyligner`: Word-alignment software based on SBI

MOLTES GRÀCIES PER L'ATENCIÓ!

THANK YOU VERY MUCH FOR YOUR ATENTION!

This work has been partially funded by:

- the Spanish government through projects TIN2009-14009-C02-01 and TIN2012-32615,
- the European Union through project Abu-MaTran (PIAP-GA-2012-324414), and
- the Kazakh Government through project UA KAZAKHUNIVERSITY1-15I.