Using Alignment Templates to Infer Shallow-Transfer Machine Translation Rules

Felipe Sánchez-Martínez[1] Hermann Ney[2]

[1]Transducens Group – Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant
E-03071 Alacant, Spain
fsanchez@dlsi.ua.es

[2]Lehrstuhl für Informatik VI – Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany
ney@informatik.rwth-aachen.de

FinTAL, 5th International Conference on Natural Language Processing
Turku/Åbo, Finland
August 23rd, 2006

Outline

Outline

Motivation

- Building rule-based MT systems: considerable human effort is needed to code transfer rules
- Transfer rules are used:
  - to produce grammatically correct translations in the target language (TL)
  - to perform some lexical changes, such as preposition changes
  - to introduce auxiliary verbs when needed
  - etc.

Goal

- To automatically learn those transformations that produce correct translations in the TL
- How? Converting alignment templates (ATs) into transfer rules to be used within a shallow-transfer MT system

Outline

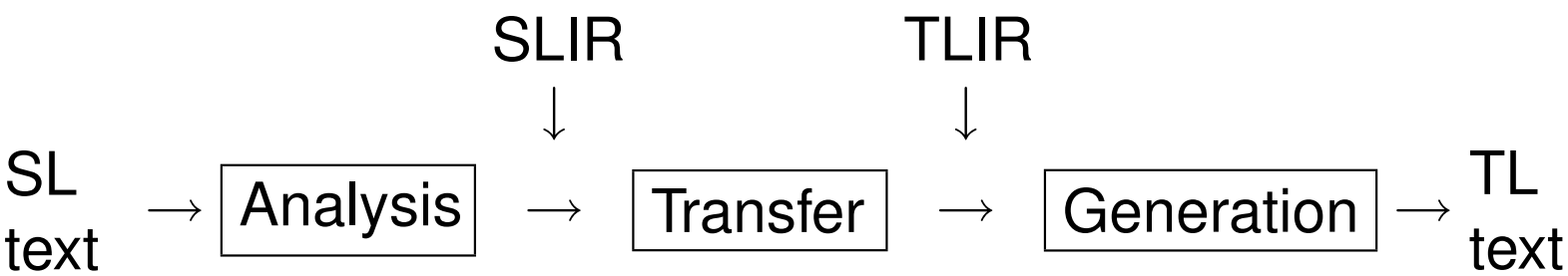
Alignment templates

- Introduced in the statistical machine translation framework as a feature function (Och and Ney 2004)
- ATs are learned in a 3-stage procedure:
  - 1^{st}: Compute word alignments
  - 2^{nd}: Extract aligned phrase pairs (*translation units*)
  - 3^{rd}: Perform a generalization over the extracted phrases using word classes
- Word classes can be automatically extracted or manually defined
- AT $z = (S_n, T_m, A)$
  - S_n: sequence of n SL word classes
  - T_m: sequence of m TL word classes
  - A: alignment information

Outline

Indirect rule-based MT

- Source language (SL) text is analyzed and converted into an intermediate representation (IR), transformations are applied and, finally, target language (TL) text is generated



- For shallow-transfer MT the IR is usually based on lemma and part-of-speech information for each word
  - *The green houses* ⟶ *the*–`(art)` *green*–`(adj)` *house*–`(noun,pl)`

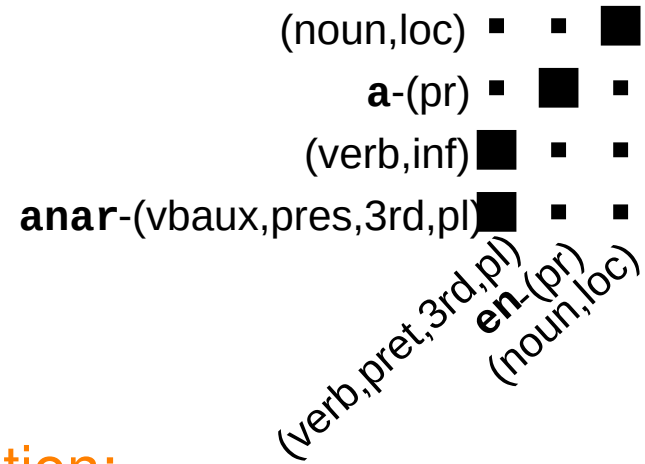
Extracting ATs for shallow-transfer MT

- Linguistic information needed:
  - closed lexical categories: Categories that cannot grow by adding new words to the dictionaries: prepositions, articles, pronouns, auxiliary verbs, ...
  - dominant categories: A dominant category is a lexical category which usually propagates its inflection information (such as gender or number) to neighboring lexical categories
    - `noun` propagates their gender and number to the `article` and the `adjective`
- The part-of-speech is used as the word class for each word, but *close words* have their own single class

Alignment template example 1

Bilingual phrase:

Alignment template:



Spanish intermediate representation:

vivieron en Alicante[1] ⟶ *vivir*-(verb,pret,3rd,pl)
en-(pr) *Alicante*-(noun,loc)

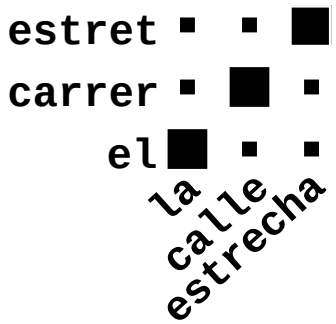
Catalan intermediate representation:

van viure a Alacant ⟶ **anar**-(vbaux,pres,3rd,pl)
viure-(verb,inf) **a**-(pr)
Alacant-(noun,loc)

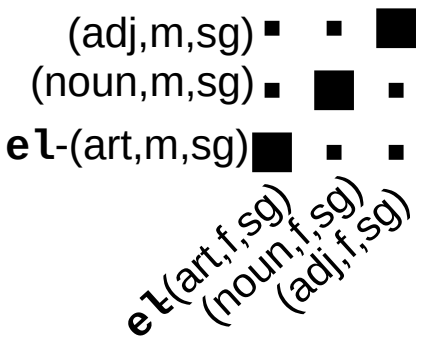
[1] Translate into English as *They lived in Alicante*

Alignment template example 2

Bilingual phrase:



Alignment template:



Spanish intermediate representation:

la calle estrecha[2] $\longrightarrow$ **el**$-$`(art,f,sg)` *calle*$-$`(noun,f,sg)`
estrecho$-$`(adj,f,sg)`

Catalan intermediate representation:

el carrer estret $\longrightarrow$ **el**$-$`(art,m,sg)` *carrer*$-$`(noun,m,sg)`
estret$-$`(adj,m,sg)`

[2] Translated into English as *The narrow street*

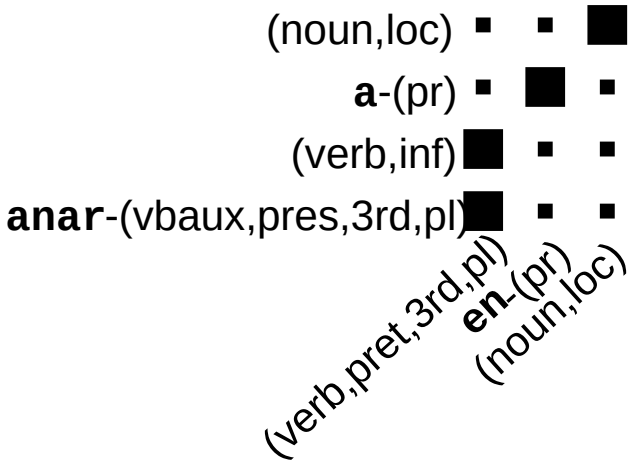
Filtering the alignment templates

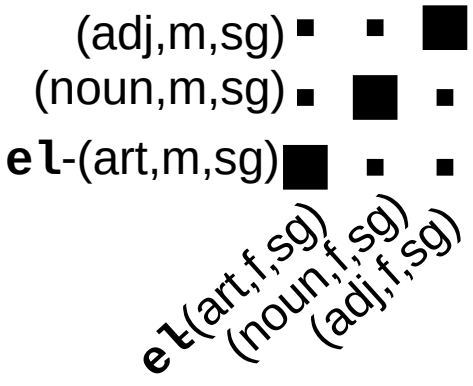
- Not all extracted ATs are applicable
- It must be ensured that the number of open categories are the same in both languages
  - It has no sense to delete for example an `adjective` in the TL
  - If a category is introduced the MT will not have any information about the lemma to be used

Alignment templates matching

- The SL part of the AT must match exactly the SL text segment to which the AT will be applied
- The inflection information provided by the bilingual dictionary for those words whose lexical category is in the dominant categories set cannot be changed by the AT

Alignment templates application





- Open words are translated by looking them up in a bilingual dictionary
- Their part-of-speech and inflection information is taken from the TL part of the AT
- SL closed words are not translated
- TL closed words are printed out as they appear in the AT
- Alignment information is used to place words in their correct order

Alignment templates application: Example

Spanish (input): *permanecieron en Alemania*[3] $\longrightarrow$

permanecer–(verb,pret,3rd,pl) **en**–(pr)

Alemania–(noun,loc)

Catalan (output): **anar**–(vbaux,pres,3rd,pl)

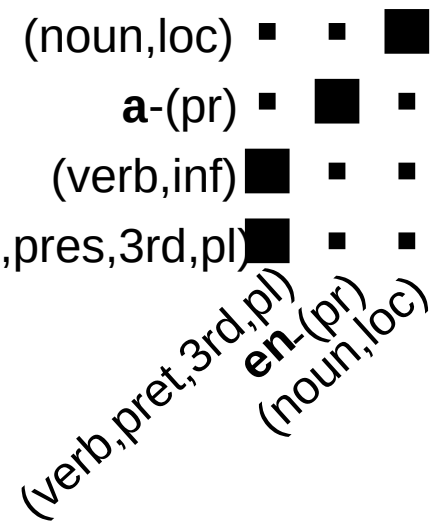
romandre–(verb,inf) **a**–(pr)

Alemanya–(noun,loc) $\longrightarrow$

van romandre a Alemanya

Word-for-word translation:

romangueren en Alemanya



[3] Translated into English as *They remained in Germany*

Outline

Experimental setup

- Using the Spanish-Catalan shallow-transfer MT system interNOSTRUM (`www.internostrum.com`)
- The transfer module is replaced by the transfer module that applies the ATs
- The performance of the system is compared to:
  - word-for-word translation (when no transfer rules are applied)
  - the original MT system using hand-coded transfer rules

Corpora

- Corpus used for training:

Language	Sentences	Running words	Vocab. size
Spanish (training)	400 000	7 480 909	157 841
Catalan (training)	400 000	7 285 133	155 446
Spanish (ATs ext.)	15 000	288 084	31 409
Catalan (ATs ext.)	15 000	296 409	30 228

- Corpus used for evaluation (only one reference translation):

Language	Sentences	Running words	Vocab. size
Spanish	1 498	32 092	7 473
Catalan	1 498	31 468	7 088

Selecting the alignment template to apply

- Two different approximations:
  - to apply the most frequent AT that matches, and
  - to apply the longest AT that matches
- In both criteria infrequent ATs are discarded according to their frequency counts
- The second criterion is suitable for a left-to-right longest-match implementation that speeds up the translation task

Results

- Spanish→Catalan translation:

MT setup	WER	PER	NIST	BLEU
word-for-word	29.41	26.99	10.07	53.07
longest AT	24.63	22.86	10.75	59.41
most frequent AT	24.50	22.70	10.77	59.75
hand-coded rules	22.94	21.05	10.88	62.50

- Catalan→Spanish translation:

MT setup	WER	PER	NIST	BLEU
word-for-word	30.01	27.46	9.76	52.59
longest AT	25.32	23.25	10.51	57.69
most frequent AT	25.90	23.78	10.44	56.66
hand-coded rules	23.77	22.19	10.53	60.23

Outline

Concluding remarks

- The use of the ATs within a shallow-transfer MT is feasible
- Some linguistic knowledge has been used; however, the linguistic information used can be easily provided
- There is a considerable improvement of the MT quality compare to word-for-word translation
- Results for both selection criteria are comparable
- Relative improvement about 70% for Spanish→Catalan, and about 60 % for Catalan→Spanish

Future work

- Study why the improvement in the MT quality is higher when translating from Spanish to Catalan
- Mix both selecting criteria into a single one
  - Use of a log-linear combination
- Avoid the use of dominant categories
- Generate transfer rules for the open-source MT engine Apertium (`http://apertium.org`)
- Try with less related language pairs like Catalan–English

Using Alignment Templates to Infer Shallow-Transfer Machine Translation Rules

Felipe Sánchez-Martínez[1] Hermann Ney[2]

[1]Transducens Group – Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant
E-03071 Alacant, Spain
fsanchez@dlsi.ua.es

[2]Lehrstuhl für Informatik VI – Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany
ney@informatik.rwth-aachen.de

FinTAL, 5th International Conference on Natural Language Processing
Turku/Åbo, Finland
August 23rd, 2006