Using unsupervised corpus-based methods to build rule-based machine translation systems

Felipe Sánchez Martínez

fsanchez@dlsi.ua.es

Ph.D. thesis
supervised by

Mikel L. Forcada

Juan Antonio Pérez Ortiz

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos

30th June 2008

Outline

1. Motivation & goal
2. Part-of-speech taggers for machine translation
   - Part-of-speech tagging
   - MT-oriented hidden Markov model training
3. Pruning of disambiguation paths
   - Disadvantages of the MT-oriented method
   - Pruning method
4. Part-of-speech tag clustering
   - Best HMM topology for taggers used in MT
   - Bottom-up agglomerative clustering
5. Automatic inference of transfer rules
   - Alignment templates for shallow-transfer machine translation
   - Generation of Apertium transfer rules
6. Concluding remarks

Outline

Motivation

- Experience in the development of shallow-transfer MT systems
  - interNOSTRUM Spanish↔Catalan
  - Traductor Universia Spanish↔Portuguese
  - Apertium Several language pairs available
- Huge human effort to code all the linguistic resources
- Resources usually needed by shallow-transfer MT systems
  - Monolingual dictionaries
  - Part-of speech (PoS) taggers
  - Bilingual dictionaries
  - Structural transfer rules

Goal

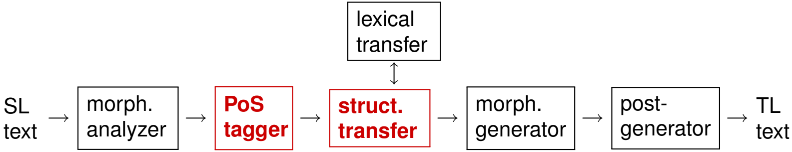
Goal:

- To reduce the human effort
- Using corpus-based methods
- In an unsupervised way

Focus on:

- the PoS taggers used in the analysis phase
- the set of shallow structural transfer rules used in translation

$\Rightarrow$ Benefiting from the rest of resources $\Leftarrow$



http://apertium.org

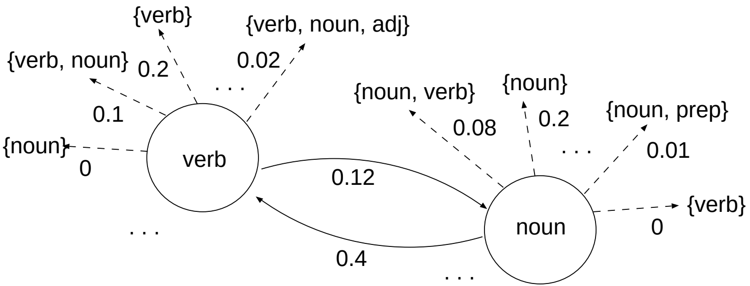
Outline

1. Motivation & goal
2. Part-of-speech taggers for machine translation
   - Part-of-speech tagging
   - MT-oriented hidden Markov model training
3. Pruning of disambiguation paths
   - Disadvantages of the MT-oriented method
   - Pruning method
4. Part-of-speech tag clustering
   - Best HMM topology for taggers used in MT
   - Bottom-up agglomerative clustering
5. Automatic inference of transfer rules
   - Alignment templates for shallow-transfer machine translation
   - Generation of Apertium transfer rules
6. Concluding remarks

Part-of-speech tagging /1

Problem: Selecting the correct PoS tag for those words with more than one (ambiguous words)
⇒ *Hidden Markov models* (HMM) are one of the standard statistical solutions

- Each HMM state corresponds to a different PoS tag
- Each input word is replaced by its corresponding ambiguity class



Part-of-speech tagging /2

In MT PoS tagging becomes crucial:

- Translation may differ from one PoS tag to another

English	PoS	Spanish
book	`noun`	*libro*
	`verb`	*reservar*

- Structural transformations may be applied (or not) for some PoS tag

English	PoS	Spanish	
the green house	*green*–`adj`	*la casa verde*	reordering ←rule applied
	green–`noun`	* *el césped casa*	

General-purpose HMM training methods

General-purpose HMM training methods:

- Supervised (hand-tagged corpora available):
  - Maximum-likelihood estimate (MLE)
- Unsupervised (only untagged corpora available):
  - Baum-Welch (expectation-maximization, EM)

Main features:

- Only use information from the language being tagged
- Independent of the natural language processing application
- To get high tagging accuracy supervised resources (hand-tagged corpora) must be built

MT-oriented HMM training method

- PoS tagging is just an intermediate task for the whole translation procedure
- Good translation performance, rather than PoS tagging accuracy, becomes the real objective

> Idea: As the goal is to get good translations into TL, let a TL model decide whether a given "construction" in the TL is good or not

MT-oriented HMM training method: overview /1

lexical transfer

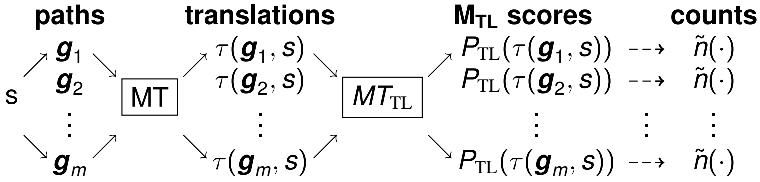
↕

SL text → morph. analyzer → **PoS tagger** → struct. transfer → morph. generator → post-generator → TL text

- Unsupervised training
- Resources required:
  - an SL untagged text automatically obtained from an SL raw corpus
  - the other modules of the MT system following the PoS tagger
  - a TL model trained from a raw TL corpus

MT-oriented HMM training method: overview /2

- Procedure:
  1. SL corpus is segmented
  2. All possible disambiguations of each segment are translated into TL
  3. A TL model is used to score each translation
  4. HMM parameters are computed according to the likelihood of the corresponding translations into TL



$\Rightarrow$ The resulting tagger is tuned to the translation fluency $\Leftarrow$

Example: English→Spanish

- SL segment (English):
  - He-`prn` rocks-`noun|verb` the-`art` table-`noun|verb`

Example: English→Spanish

- SL segment (English):
  - He-`prn` rocks-`noun|verb` the-`art` table-`noun|verb`

- Possible translations (Spanish) according to each disambiguation and their normalized likelihoods according to a TL model:

• Él-`prn` mece-`verb` la-`art` mesa-`noun`	0.75
• Él-`prn` mece-`verb` la-`art` presenta-`verb`	0.15
• Él-`prn` rocas-`noun` la-`art` mesa-`noun`	0.06
• Él-`prn` rocas-`noun` la-`art` presenta-`verb`	+ 0.04
	1.00

Example: English→Spanish

- SL segment (English):
  - He-`prn` rocks-`noun|verb` the-`art` table-`noun|verb`

- Possible translations (Spanish) according to each disambiguation and their normalized likelihoods according to a TL model:

• Él-`prn` mece-`verb` la-`art` mesa-`noun`	0.75
• Él-`prn` mece-`verb` la-`art` presenta-`verb`	0.15
• Él-`prn` rocas-`noun` la-`art` mesa-`noun`	0.06
• Él-`prn` rocas-`noun` la-`art` presenta-`verb`	+ 0.04
	1.00

- The HMM parameters involved in these 4 disambiguations are updated according to their likelihoods in the TL

Experiments /1

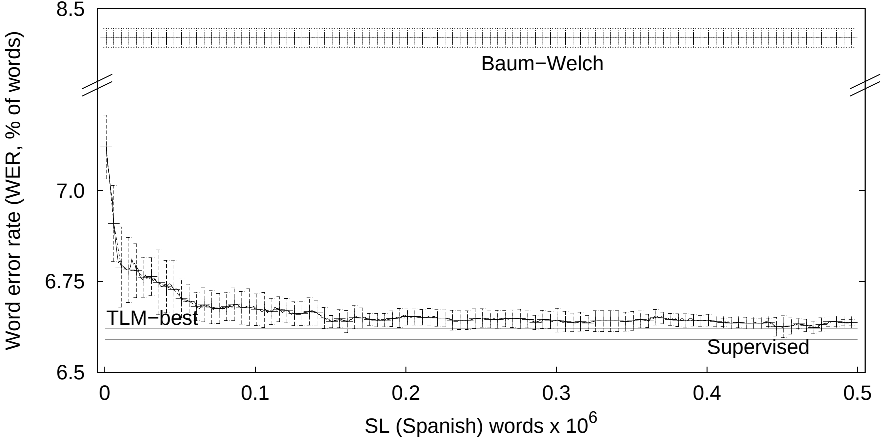
- Task: training PoS tagger for Spanish, French and Occitan to be used in MT into Catalan
- TL model: trigram language model trained from a Catalan corpus with $\approx 2 \cdot 10^6$ words
- Experiments conducted with
  - 5 disjoint corpora with $0.5 \cdot 10^6$ words for Spanish
  - 5 disjoint corpora with $0.5 \cdot 10^6$ words for French
  - Only one corpus with $0.3 \cdot 10^6$ words for Occitan

Experiments /2

- Reference results:
  - Baum-Welch expectation maximization on $10 \cdot 10^6$ words corpora
  - Supervised: MLE from a hand-tagged corpus $\approx 21.5 \cdot 10^3$ words (only for Spanish)
  - TLM-best: when a TL model is used at translation time to select always the most likely translation
    - approximate indication of the best results the MT-oriented method could achieve

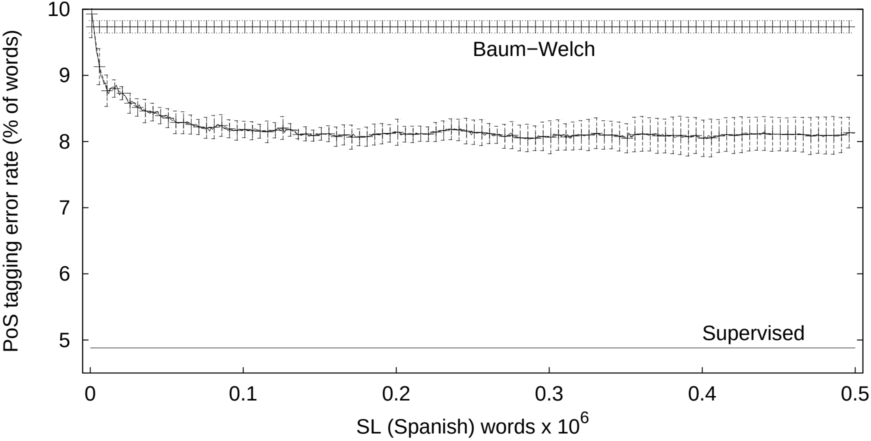
Some results: Spanish→Catalan /1

Mean and std. dev. of the translation performance, WER (% of words)



Some results: Spanish→Catalan /2

Mean and std. dev. of the PoS tagging error rate (% of words)



Some results: Spanish→Catalan /3

Why are the translation performances for the supervised and the MT-oriented method comparable, but no the PoS tagging error rates?

- TL information does not discriminate among the SL analyses of a segment leading to the same translation

French	**PoS**	**Spanish**
la ville	*la*-`art` *la*-`prn`	*la ciudad*

- *Free-ride*: phenomenon by which choosing the incorrect interpretation for an ambiguous word does not result in a translation error

Outline

Disadvantages of the MT-oriented method

- The number of possible disambiguations to translate grows exponentially with segment length
- Translation is the most time-consuming task
- Goal: To overcome this problem
- How: Pruning unlikely disambiguation paths by using *a priori* knowledge

Pruning method /1

- Based on an initial model of SL tags

> Assumption: Any reasonable model of SL tags may be useful to choose a subset of possible disambiguation paths so that the correct one is in that subset

- The model used for pruning can be updated dynamically during training

Pruning method /2

1. The *a priori* likelihood of each possible disambiguation path of SL segment *s* is calculated using the pruning model
2. The set of disambiguation paths to take into account is determined by using a mass probability threshold ρ
    - Only the minimum number of paths to reach the mass probability threshold ρ are taken into account

Example (English→Spanish)

- SL segment (English):
  - He-`prn` rocks-`noun|verb` the-`art` table-`noun|verb`

Example (English→Spanish)

- SL segment (English):
  - He-`prn` rocks-`noun|verb` the-`art` table-`noun|verb`

- Normalized *a priori* likelihoods:

$$
\begin{array}{ll}
\boldsymbol{g}_1 = (\mathtt{prn}, \mathtt{verb}, \mathtt{art}, \mathtt{noun}) & \,0.69 \\
\boldsymbol{g}_2 = (\mathtt{prn}, \mathtt{verb}, \mathtt{art}, \mathtt{verb}) & \,0.14 \\
\boldsymbol{g}_3 = (\mathtt{prn}, \mathtt{noun}, \mathtt{art}, \mathtt{noun}) & \,0.10 \\
\boldsymbol{g}_4 = (\mathtt{prn}, \mathtt{noun}, \mathtt{art}, \mathtt{verb}) & +\,0.07 \\
\hline
 & \,1.00
\end{array}
$$

Example (English→Spanish)

- SL segment (English):
    - He-`prn` rocks-`noun|verb` the-`art` table-`noun|verb`

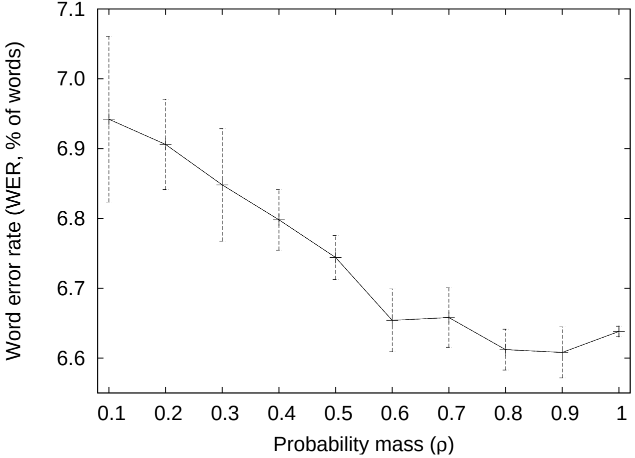
- Normalized *a priori* likelihoods:

$$
\begin{array}{ll}
\boldsymbol{g}_1 = (\texttt{prn}, \texttt{verb}, \texttt{art}, \texttt{noun}) & \,0.69 \\
\boldsymbol{g}_2 = (\texttt{prn}, \texttt{verb}, \texttt{art}, \texttt{verb}) & \,0.14 \\
\boldsymbol{g}_3 = (\texttt{prn}, \texttt{noun}, \texttt{art}, \texttt{noun}) & \,0.10 \\
\boldsymbol{g}_4 = (\texttt{prn}, \texttt{noun}, \texttt{art}, \texttt{verb}) & +\,0.07 \\
\hline
 & \,1.00
\end{array}
$$

- With $\rho = 0.8$, $\boldsymbol{g}_3$ and $\boldsymbol{g}_4$ are discarded because $0.69 + 0.14 \geq 0.8$

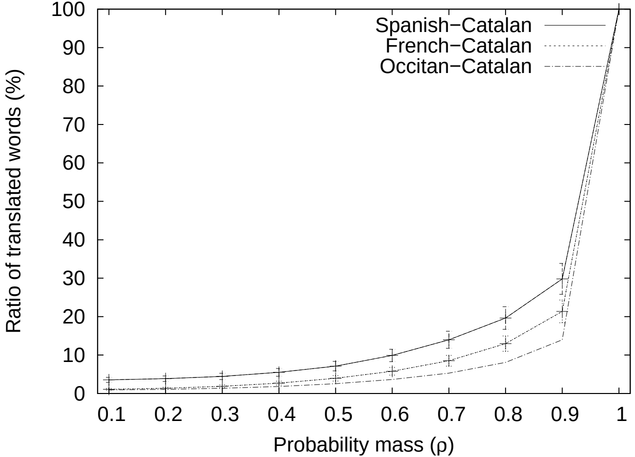
Some results: Spanish→Catalan

Mean and std. dev. of the translation performance, WER (% of words)



Some results

Ratio of translated words



Outline

Best HMM topology for taggers used in MT

- Large tagsets (set of PoS tags) for richly-inflected languages
  - fine PoS tags convey lot of information
    e.g. `verb.pret.3rd.pl`, `noun.m.sg`
- A reduced tagset manually defined following linguistic guidelines is usually used
  - Maps fine tags into coarse ones
  - Should allow for better parameter estimation
- Goal: To automatically determine the set of states to be used
  - Avoid the human intervention in defining the tagset

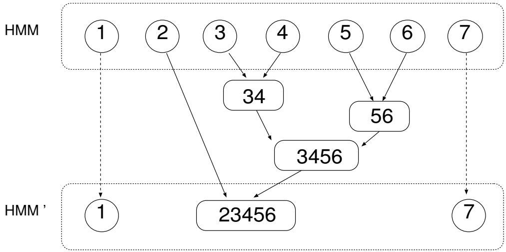
⇒ *Model merging* approach (Stolcke and Omohundro, 1994) cannot be applied using untagged corpora

Bottom-up agglomerative clustering

1. Place each object in its own cluster (singleton)
2. Iteratively compare all pairs of clusters and choose the two closest clusters according to a distance measure
   - If the distance between the selected clusters is below a certain threshold, merge both clusters
   - Otherwise, stop

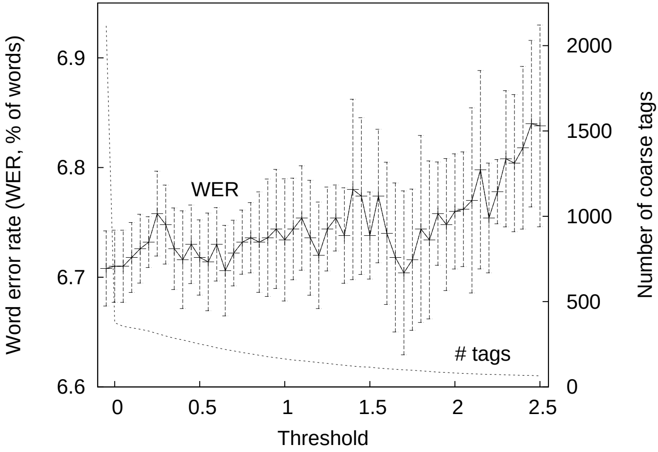
Clustering of PoS tag

- First model trained using the large tagset via the MT-oriented method
- Distance between cluster based on the state-to-state transition probabilities
- An additional constraint ensures that it is possible to restore the information about the fine tag from the coarse one



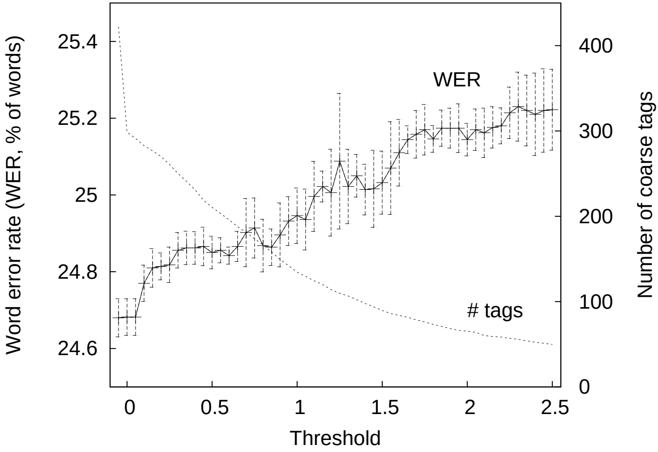
Some results: Spanish→Catalan

Mean and std. dev. of the translation performance, WER



Some results: French→Catalan

Mean and std. dev. of the translation performance, WER



Outline

1. Motivation & goal
2. Part-of-speech taggers for machine translation
   - Part-of-speech tagging
   - MT-oriented hidden Markov model training
3. Pruning of disambiguation paths
   - Disadvantages of the MT-oriented method
   - Pruning method
4. Part-of-speech tag clustering
   - Best HMM topology for taggers used in MT
   - Bottom-up agglomerative clustering
5. Automatic inference of transfer rules
   - Alignment templates for shallow-transfer machine translation
   - Generation of Apertium transfer rules
6. Concluding remarks

Automatic inference of transfer rules

Goal:

- To automatically learn those transformations that produce correct translations in the TL

How:

- Adapting the alignment templates (ATs) already used in statistical MT to the shallow-transfer approach
  - AT $z = (S_n, T_m, G)$
    - S_n: sequence of n SL word classes
    - T_m: sequence of m TL word classes
    - G: alignment information

AT for shallow-transfer MT: overview /1

- Resources required:
  - A SL–TL parallel corpus
  - The morphological analyzers and PoS taggers of the MT system
  - The bilingual dictionary of the MT system

- Procedure:
  1. Analyze both sides of the training corpus
  2. Compute word alignments
  3. Extract bilingual phrase pairs and derive ATs from them
  4. Generate shallow-transfer rules

AT for shallow-transfer MT: overview /2

- Word class: part-of-speech (including all the inflection information)
  - Exception: lexicalized words are placed in single-word classes
- Lexicalized categories: categories that are known to be involved in lexical changes, such as prepositions
  - the method can learn not only syntactic changes

AT for shallow-transfer MT: overview /3

- ATs are extended with a set R of restrictions over the TL inflection information of non-lexicalized words
  - AT $z = (S_n, T_m, G, R)$

AT for shallow-transfer MT: overview /3

- ATs are extended with a set R of restrictions over the TL inflection information of non-lexicalized words
  - AT $z = (S_n, T_m, G, R)$
- Restrictions are derived from the bilingual dictionary
  - Bilingual entry that does not change inflection information

```
<e><p>
  <l>castigo<s n="noun"/></l>
  <r>càstig<s n="noun"/></r>
</p></e>
```

R: w=`noun.*`

  - Bilingual entry that does change inflection information

```
<e><p>
  <l>calle<s n="noun"/><s n="f"/></l>
  <r>carrer<s n="noun"/><s n="m"/></r>
</p></e>
```

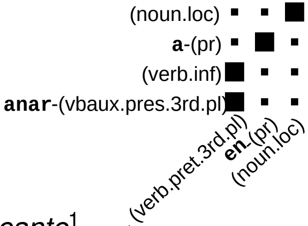
R: w=`noun.`**m**`.*`

- The bilingual dictionary is also used to discard phrase pairs that cannot be reproduced by the MT system

Alignment template example /1

Bilingual phrase:

Alignment template:



Spanish analysis: *vivieron en Alicante*[1] $\longrightarrow$
vivir-`(verb.pret.3rd.pl)` **en**-`(pr)`
Alicante-`(noun.loc)`

Catalan analysis: *van viure a Alacant* $\longrightarrow$
anar-`(vbaux.pres.3rd.pl)` *viure*-`(verb.inf)`
a-`(pr)` *Alacant*-`(noun.loc)`

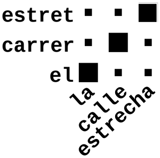
Restrictions: w_2 =`verb.*`, w_4=`noun.*`

---

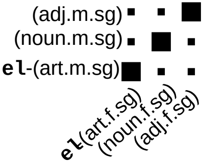
[1] Translated into English as *They lived in Alicante*

Alignment template example /2

Bilingual phrase:



Alignment template:



Spanish analysis: *la calle estrecha*[2] $\longrightarrow$ **el**-(art.f.sg) *calle*-(noun.f.sg) *estrecho*-(adj.f.sg)

Catalan analysis: *el carrer estret* $\longrightarrow$ **el**-(art.m.sg) *carrer*-(noun.m.sg) *estret*-(adj.m.sg)

Restrictions: w_2 =noun.m.$\star$, w_3=adj.$\star$

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

Generation of Apertium transfer rules

Procedure:

1. Discard useless AT
2. Select the AT to use according to their frequency
3. For all ATs with the same SL part a rule is generated

Rule generation:

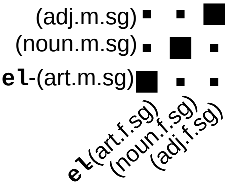
- The rule matches the SL part all ATs have in common
- In decreasing order of AT frequency counts code is generated to
  - test the restrictions *R* over the TL inflection information
  - if they hold, apply the AT and stop rule execution
- code that translates word-for-word is added
  - it is executed only if none of the AT were *applicable*

AT applicability test /1

Restrictions *R* are tested by looking at the bilingual dictionary

Example:

- *R*: w_2 =**noun.m**`.*`, w_3=`adj.*`
- Input string (Spanish): *la señal roja* $\longrightarrow$ **el**–`(art.f.sg)` *señal*–`(noun.f.sg)` *rojo*–`(adj.f.sg)`
- Translation of non-lexicalized words:
  - *señal*–`(noun.f.sg)`→*senyal*–`(`**noun.m**`.sg)`
  - *rojo*–`(adj.f.sg)`→*vermell*–`(adj.f.sg)`
- Restriction holds, AT can be applied

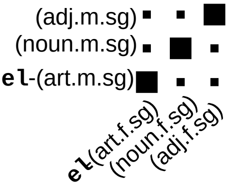


AT applicability test /2

Restrictions *R* are tested by looking at the bilingual dictionary

Example:

- *R*: w_2 =**noun.m**.$\star$, w_3=`adj.`$\star$
- Input string (Spanish): *la silla blanca* $\longrightarrow$ **el**-`(art.f.sg)` *silla*-`(noun.f.sg)` *blanco*-`(adj.f.sg)`
- Translation of non-lexicalized words:
  - *silla*-`(noun.f.sg)`$\rightarrow$*cadira*-(**noun.f**.`sg`)
  - *blanco*-`(adj.f.sg)`$\rightarrow$*blanc*-`(adj.f.sg)`
- Restriction does not hold, AT cannot be applied



Alignment templates application, an example

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

permanecer-`(verb.pret.3rd.pl)` **en**-`(pr)`
Alemania-`(noun.loc)`

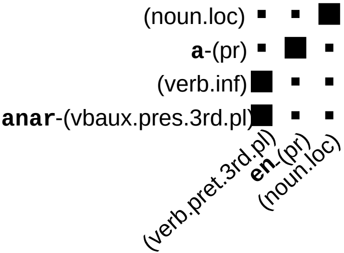
Catalan (output): **anar**-`(vbaux.pres.3rd.pl)`

romandre-`(verb.inf)` **a**-`(pr)`
Alemanya-`(noun.loc)` $\longrightarrow$
van romandre a Alemanya

Word-for-word translation:

*romangueren *en Alemanya*

R: w_1 =`verb.*`, w_3=`noun.*`



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

Experiments

- Task: Inference of shallow-transfer rules for Spanish↔Catalan, Spanish↔Galician and Spanish→Portuguese
- ≈ 8 lexicalized categories
- Two different training corpora:
  - One with $2 \cdot 10^6$ words
  - Another with only $0.5 \cdot 10^6$ words
- Two different evaluation corpora:

  post-edit reference translation is a post-edited version of the MT performed using hand-coded transfer rules

  parallel text to translate and reference translation comes from a parallel corpus analogous to the one used for training

Some results

Spanish→Catalan, WER ± 95% confidence interval

Training	Test	Word-for-word	AT transfer	Hand
$2 \cdot 10^6$	post-edit	12.6 ± 0.9	8.7 ± 0.7	6.7 ± 0.7
	parallel	26.4 ± 1.2	20.3 ± 1.1	20.7 ± 1.0
$0.5 \cdot 10^6$	post-edit	12.6 ± 0.9	9.9 ± 0.7	6.7 ± 0.7
	parallel	26.4 ± 1.2	21.4 ± 1.1	20.7 ± 1.0

Spanish→Portuguese, WER ± 95% confidence interval

Training	Test	Word-for-word	AT transfer	Hand
$2 \cdot 10^6$	post-edit	11.9 ± 0.8	12.1 ± 0.9	7.0 ± 0.7
	parallel	47.9 ± 1.7	46.5 ± 1.7	47.6 ± 1.8
$0.5 \cdot 10^6$	post-edit	11.9 ± 0.8	12.1 ± 0.9	7.0 ± 0.7
	parallel	47.9 ± 1.7	47.4 ± 1.7	47.6 ± 1.8

Some results

Why such a large difference between Spanish→Catalan and Spanish→Portuguese?

- Because of how training corpora have been built
  - Spanish→Catalan, by translating one language into another (newspaper *El Periódico de Catalunya*)
    - 22% of discarded ATs
  - Spanish→Portuguese, by translating from a third language (*JRC-ACQUIS* parallel corpus)
    - 53% of discarded ATs

Outline

Concluding remarks /1

Steps towards more efficient development of RBMT systems

- A new method to train PoS tagger to be used in MT
  - focuses on the task in which it will be used
  - uses TL information without using parallel corpora
  - benefits from information in the rest of modules
  - using *a priori* knowledge saves around 80% of the translations to perform while training
  - better translation quality than tagging accuracy
- PoS tags clustering
  - has not provided the expected results, but
  - may be useful if the number of states is crucial

Concluding remarks /2

- A method to infer shallow-transfer rules from parallel corpora
  - extends the definition of alignment template
  - small amount of information provided by human is used
  - the process followed to build the parallel corpus deserves special attention
  - inferred rules are human-readable
  - they can coexist with hand-coded rules

Concluding remarks /3

- Open-source software
  - Can be downloaded from `sf.net/projects/apertium`
    - Packages `apertium-tagger-training-tools` and `apertium-transfer-tools`
  - Ensures reproducibility
  - Allows other researchers to improve them
  - Eases the development of new language pairs for Apertium
  - `apertium-tagger-training-tools` is being used by Prompsit Language Enginnering S.L.

Future research lines

This thesis opens several research lines:

- the use of TL information to train other statistical models that run on the SL
- the use of more than one TL (triangulation)
- the use of a TL model of different nature
- linguistically-driven extraction of bilingual phrases
- a more flexible way to use lexicalized categories
- a bootstrapping method to learn both the PoS tagger and the set of transfer rules cooperatively
- ...

Acknowledgments

- Spanish Ministry of Education and Science, and European Social Fund; research grant BES-2004-4711
- Spanish Ministry of Industry, Commerce and Tourism; projects TIC2003-08681-C02-01, FIT340101-2004-3 and FIT-350401-2006-5
- Autonomous Government of Catalonia; project *Traducció automàtica de codi obert per al català*
- Spanish Ministry of Education and Science; project TIN2006-15071-C03-01

⇒Thank you very much for your attention⇐