Cooperative unsupervised training of the part-of-speech taggers in a bidirectional machine translation system

Felipe Sánchez-Martínez, Juan Antonio Pérez-Ortiz, Mikel L. Forcada
Departament de Llenguatges i Sistemes Informàtics
Universitat d’Alacant
E-03071 Alacant, Spain

{fsanchez,japerez,mlf}@dlsi.ua.es

*Funded by the Spanish Government through grants TIC2003-08681-C02-01 and BES-2004-4711
Contents

• Introduction
• Part-of-speech ambiguities in machine translation
• Part-of-speech tagging with HMM
• Target-language based training of HMM-based taggers
• Cooperative learning of HMM
• Experiments
• Discussion
• Future work
Introduction

**Part-of-speech (PoS) tagging:** To determine the lexical category or PoS of each word that appears in a text

**Ambiguous word:** Word with more than one possible lexical category (PoS)

<table>
<thead>
<tr>
<th>Lemma</th>
<th>PoS</th>
</tr>
</thead>
<tbody>
<tr>
<td>book</td>
<td>noun</td>
</tr>
<tr>
<td>book</td>
<td>verb</td>
</tr>
</tbody>
</table>

Ambiguities are usually solved by looking at the context
PoS ambiguities in machine translation (I)

**Indirect MT system:** Source language (SL) text is analysed and transformed into an abstract intermediate representation, transformations are applied and, finally, target language (TL) text is generated.

- Analysis module usually includes a PoS tagger
PoS ambiguities in machine translation (II)

Mistranslation due to wrong PoS tagging

- Translation differs from one PoS to another:

<table>
<thead>
<tr>
<th>Spanish</th>
<th>PoS</th>
<th>Translation into Catalan</th>
</tr>
</thead>
<tbody>
<tr>
<td>para</td>
<td>preposition</td>
<td>per a (for/to)</td>
</tr>
<tr>
<td></td>
<td>verb</td>
<td>para (stop)</td>
</tr>
</tbody>
</table>
PoS ambiguities in machine translation (II)

Mistranslation due to wrong PoS tagging

- Translation differs from one PoS to another:

<table>
<thead>
<tr>
<th>Spanish</th>
<th>PoS</th>
<th>Translation into Catalan</th>
</tr>
</thead>
<tbody>
<tr>
<td>para</td>
<td>preposition verb</td>
<td>per a (for/to)</td>
</tr>
<tr>
<td>para</td>
<td>verb</td>
<td>para (stop)</td>
</tr>
</tbody>
</table>

- Some transformation is applied (or not) for some PoS:

<table>
<thead>
<tr>
<th>Spanish</th>
<th>PoS</th>
<th>Translation into Catalan</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>las calles</td>
<td>la (article)</td>
<td>els carrers (the streets)</td>
<td>gender</td>
</tr>
<tr>
<td>la (pronoun)</td>
<td>* la (pronoun)</td>
<td>* les carrers (them streets)</td>
<td>←agreement rule applied</td>
</tr>
</tbody>
</table>
PoS tagging with HMM (I)

Use of a hidden Markov model (HMM):

- Adopting a reduced tag set (grouping the finer tags delivered by the morphological analyser)
- Each HMM state corresponds to a different PoS tag
- Each input word is replaced by its corresponding ambiguity class
Cooperative unsupervised training of the part-of-speech taggers in a bidirectional machine translation system

PoS tagging with HMM (II)

Estimating proper HMM parameters

Training

\(\begin{cases}
\text{supervised} \\
\text{unsupervised}
\end{cases}\)
PoS tagging with HMM (II)

Estimating proper HMM parameters

Training $\{\text{supervised, unsupervised}\}$

- *tagged corpus*
- *untagged corpus*
PoS tagging with HMM (II)

Estimating proper HMM parameters

Training

\[
\begin{align*}
\text{tagged corpus} 
& \quad \text{supervised} \\
\text{untagged corpus} 
& \quad \text{unsupervised} \\
\end{align*}
\]

Baum-Welch

New idea: Use of TL information
Target-language based training of HMM-based taggers (I)

- Transition probabilities

\[ a_{\gamma_i \gamma_j} = \frac{\tilde{n}(\gamma_i \gamma_j)}{\sum_{\gamma_k \in \Gamma} \tilde{n}(\gamma_i \gamma_k)} \]

- Emission probabilities

\[ b_{\gamma_i \sigma} = \frac{\tilde{n}(\sigma, \gamma_i)}{\sum_{\sigma' : \gamma_i \in \sigma'} \tilde{n}(\sigma', \gamma_i)} \]
Target-language based training of HMM-based taggers (II)

SL text $\rightarrow$ Segmentation

$\rightarrow$ segment $s_1$
$\rightarrow$ segment $s_2$
$\rightarrow$ segment $s_n$
Target-language based training of HMM-based taggers (II)

![Diagram](https://example.com/diagram.png)
Target-language based training of HMM-based taggers (II)
Target-language based training of HMM-based taggers (II)

Cooperative unsupervised training of the part-of-speech taggers in a bidirectional machine translation system
Target-language based training of HMM-based taggers (II)

\[
\text{SL text} \xrightarrow{} \text{Segmentation} \xrightarrow{} \text{segment } s_1 \xrightarrow{} \text{segment } s_2 \xrightarrow{} \text{segment } s_n
\]

\[
\text{disambiguations}\quad \text{translations}\quad \text{likelihoods}
\]
\[
\text{seg. } s_i \xrightarrow{} \text{path } g_1 \xrightarrow{} \text{path } g_2 \xrightarrow{} \text{path } g_m
\]
\[
\tau(g_1, s) \quad \tau(g_2, s) \quad \tau(g_m, s)
\]
\[
p_{TL}(\tau(g_1, s)) \rightarrow p(g_1 | s) \quad p_{TL}(\tau(g_2, s)) \rightarrow p(g_2 | s) \quad p_{TL}(\tau(g_m, s)) \rightarrow p(g_m | s)
\]
## Target-language based training of HMM-based taggers (III)

\[ s \equiv \{ \text{CNJ} \} \quad \{ \text{ART} \} \quad \{ \text{VB} \} \quad \{ \text{CNJ} \} \]

<table>
<thead>
<tr>
<th>( g_1 \equiv )</th>
<th>CNJ</th>
<th>ART</th>
<th>PR</th>
<th>CNJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau(g_1, s) \equiv )</td>
<td>( i ) (and)</td>
<td>( \text{la} ) (the)</td>
<td>( \text{per a} ) (for/to)</td>
<td>( \text{si} ) (if)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( g_2 \equiv )</th>
<th>CNJ</th>
<th>ART</th>
<th>VB</th>
<th>CNJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau(g_2, s) \equiv )</td>
<td>( i ) (and)</td>
<td>( \text{la} ) (the)</td>
<td>( \text{para} ) (stop)</td>
<td>( \text{si} ) (if)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( g_3 \equiv )</th>
<th>CNJ</th>
<th>PRN</th>
<th>PR</th>
<th>CNJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau(g_3, s) \equiv )</td>
<td>( i ) (and)</td>
<td>( \text{la} ) (it/her)</td>
<td>( \text{per a} ) (for/to)</td>
<td>( \text{si} ) (if)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( g_4 \equiv )</th>
<th>CNJ</th>
<th>PRN</th>
<th>VB</th>
<th>CNJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau(g_4, s) \equiv )</td>
<td>( i ) (and)</td>
<td>( \text{la} ) (it/her)</td>
<td>( \text{para} ) (stop)</td>
<td>( \text{si} ) (if)</td>
</tr>
</tbody>
</table>
Cooperative unsupervised training of the part-of-speech taggers in a bidirectional machine translation system

Target-language based training of HMM-based taggers (III)

<table>
<thead>
<tr>
<th>$s$</th>
<th>$y$</th>
<th>$la$</th>
<th>$para$</th>
<th>$si$</th>
</tr>
</thead>
<tbody>
<tr>
<td>{ CNJ }</td>
<td>{ ART PRN }</td>
<td>{ VB PR }</td>
<td>{ CNJ }</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$g_1$</th>
<th>CNJ</th>
<th>ART</th>
<th>PR</th>
<th>CNJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau(g_1, s)$</td>
<td>$i$ (and)</td>
<td>$la$ (the)</td>
<td>$per$ (for/to)</td>
<td>$si$ (if)</td>
</tr>
<tr>
<td>$p(g_1</td>
<td>s)$</td>
<td>0.0001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$g_2$</th>
<th>CNJ</th>
<th>ART</th>
<th>VB</th>
<th>CNJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau(g_2, s)$</td>
<td>$i$ (and)</td>
<td>$la$ (the)</td>
<td>$para$ (stop)</td>
<td>$si$ (if)</td>
</tr>
<tr>
<td>$p(g_2</td>
<td>s)$</td>
<td>0.4999</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$g_3$</th>
<th>CNJ</th>
<th>PRN</th>
<th>PR</th>
<th>CNJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau(g_3, s)$</td>
<td>$i$ (and)</td>
<td>$la$ (it/her)</td>
<td>$per$ (for/to)</td>
<td>$si$ (if)</td>
</tr>
<tr>
<td>$p(g_3</td>
<td>s)$</td>
<td>0.0001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$g_4$</th>
<th>CNJ</th>
<th>PRN</th>
<th>VB</th>
<th>CNJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau(g_4, s)$</td>
<td>$i$ (and)</td>
<td>$la$ (it/her)</td>
<td>$para$ (stop)</td>
<td>$si$ (if)</td>
</tr>
<tr>
<td>$p(g_4</td>
<td>s)$</td>
<td>0.4999</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Free ride:** word translated the same way independently of the tag selected
Target-language based training of HMM-based taggers (IV)

\[ p(g_i|s) \propto p(g_i|\tau(g_i, s)) p_{\text{TL}}(\tau(g_i, s)) \]

- \( p(g_i|s) \): Probability of path \( g_i \) to be the correct disambiguation of segment \( s \)
- \( p_{\text{TL}}(\tau(g_i, s)) \): Likelihood of the translation into TL of segment \( s \) according to the disambiguation given by path \( g_i \)
  - Language model based on trigrams of words
  - Hidden Markov model
  - ...
- \( p(g_i|\tau(g_i, s)) \): Contribution of the disambiguation path \( g_i \) to the translation given by \( \tau(g_i, s) \)
Cooperative learning of HMM (I)

- Use of the previous idea ...
Cooperative learning of HMM (I)

- Use of the previous idea ...
- Bidirectional MT system translating between languages $A$ and $B$
Cooperative learning of HMM (I)

- Use of the previous idea ...
- Bidirectional MT system translating between languages $A$ and $B$
- Morphological generation is not done when performing translations
Cooperative learning of HMM (I)

- Use of the previous idea ...

- Bidirectional MT system translating between languages $A$ and $B$

- Morphological generation is not done when performing translations

- Before morphological generation we have a sequence of lexical categories (tags) in the TL
Cooperative learning of HMM (I)

- Use of the previous idea ...

- Bidirectional MT system translating between languages $A$ and $B$

- Morphological generation is not done when performing translations

- Before morphological generation we have a sequence of lexical categories (tags) in the TL

- Use of such a TL model based on tags: HMM as a TL model
Cooperative learning of HMM (II)


\[
\begin{align*}
M_A[0] & \rightarrow M_B[0] \\
\vdots & \vdots \\
M_A[k] & \rightarrow M_B[k] \\
\vdots & \vdots \\
M_B[k-1] & \\
\vdots & \\
M_B[k] & \\
\end{align*}
\]
Experiments

- We used the Spanish↔Catalan MT system interNOSTRUM (www.internostrum.com)

  Language $A$: Catalan
  Language $B$: Spanish
Experiments

- We used the Spanish ↔ Catalan MT system interNOSTRUM (www.internostrum.com)
  
  Language $A$: Catalan  
  Language $B$: Spanish

- Use of various corpus sizes and three different corpora for each size
Experiments

• We used the Spanish↔Catalan MT system interNOSTRUM (www.internostrum.com)
  Language A: Catalan
  Language B: Spanish

• Use of various corpus sizes and three different corpora for each size

• Evaluation with an independent corpus for each language:
  – PoS error rate with hand-tagged corpus
  – Translation error rate with human-corrected translations

TMI, Baltimore 4–6 October, 2004
Results

Proof of two different initial models $M_B[0]$: 

- “Good” HMM: Trained from 1,000,000-word untagged SL corpus with the Baum-Welch algorithm (PoS error rate: 34.2%)

- “Bad” HMM: Equiprobable transition and emission probabilities (PoS error rate: 76.5%)
Results

Proof of two different initial models $M_B[0]$: 

- “Good” HMM: Trained from 1,000,000-word untagged SL corpus with the Baum-Welch algorithm (PoS error rate: 34.2%)
- “Bad” HMM: Equiprobable transition and emission probabilities (PoS error rate: 76.5%)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spanish</td>
<td>Catalan</td>
<td>Spanish</td>
</tr>
<tr>
<td>“good start”</td>
<td>24.9%</td>
<td>27.5%</td>
<td>6.2%</td>
</tr>
<tr>
<td>“bad start”</td>
<td>25.9%</td>
<td>26.4%</td>
<td>6.1%</td>
</tr>
<tr>
<td>Baum-Welch</td>
<td>31.7%</td>
<td>37.8%</td>
<td>8.4%</td>
</tr>
<tr>
<td>supervised</td>
<td>10.4%</td>
<td>16.5%</td>
<td>2.6%</td>
</tr>
</tbody>
</table>

TMI, Baltimore 4–6 October, 2004
Discussion

- PoS error and translation error rates lie between those produced by supervised and unsupervised methods
Discussion

• PoS error and translation error rates lie between those produced by supervised and unsupervised methods

• There is no need for good initial information to achieve good results
Discussion

- PoS error and translation error rates lie between those produced by supervised and unsupervised methods.

- There is no need for good initial information to achieve good results.

- The method described needs a relatively small amount of words compared with common corpus sizes used with the Baum-Welch algorithm.
Discussion

- PoS error and translation error rates lie between those produced by supervised and unsupervised methods
- There is no need for good initial information to achieve good results
- The method described needs a relatively small amount of words compared with common corpus sizes used with the Baum-Welch algorithm
- The training method produces PoS taggers tuned not only with SL texts, but also with TL texts and the underlying MT system
Future work

- Research on better estimates for $p(g_i | \tau(g_i, s))$
  - Estimate the HMM parameters iteratively
    Use the parameters of the previous iteration to estimate $p(g_i | \tau(g_i, s))$
Future work

- Research on better estimates for $p(g_i|\tau(g_i, s))$
  - Estimate the HMM parameters iteratively
    Use the parameters of the previous iteration to estimate $p(g_i|\tau(g_i, s))$

- Time complexity reduction
  - Use of a $k$-best Viterbi algorithm with the current parameters to calculate approximate likelihood and translate only the $k$ most promising paths
Future work

• Research on better estimates for $p(g_i | \tau(g_i, s))$
  
  – Estimate the HMM parameters iteratively
    Use the parameters of the previous iteration to estimate $p(g_i | \tau(g_i, s))$

• Time complexity reduction

  – Use of a $k$-best Viterbi algorithm with the current parameters to calculate
    approximate likelihood and translate only the $k$ most promising paths

• Better formalization

  – Different disambiguation paths from different segments can produce the
    same translation
Thank you very much for your attention !!