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 (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)

	Lemma	**PoS**
book	*book*	noun
	book	verb

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.

$$\begin{array}{c}\text{SL}\\\text{text}\end{array} \longrightarrow \boxed{\text{Analysis}} \xrightarrow{\text{SLAR}} \boxed{\text{Transformation}} \xrightarrow{\text{TLAR}} \boxed{\text{Generation}} \longrightarrow \begin{array}{c}\text{TL}\\\text{text}\end{array}$$

- 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:

Spanish	**PoS**	**Translation into Catalan**
para	preposition	*per a* (for/to)
	verb	*para* (stop)

PoS ambiguities in machine translation (II)

Mistranslation due to wrong PoS tagging

- Translation differs from one PoS to another:

Spanish	PoS	Translation into Catalan
para	preposition	*per a* (for/to)
	verb	*para* (stop)

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

Spanish	PoS	Translation into Catalan
las calles	*la* (article)	*els carrers* (the streets)
	la (pronoun)	* *les carrers* (them streets)

←gender agreement rule applied

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

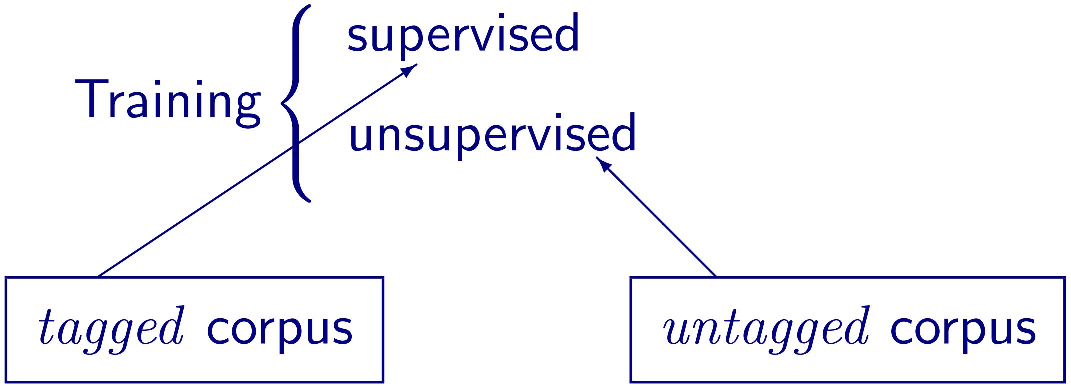
PoS tagging with HMM (II)

Estimating proper HMM parameters

$$\text{Training} \begin{cases} \text{supervised} \\ \text{unsupervised} \end{cases}$$

PoS tagging with HMM (II)

Estimating proper HMM parameters



PoS tagging with HMM (II)

Estimating proper HMM parameters

$$\text{Training} \begin{cases} \text{supervised} \\ \text{unsupervised} \begin{cases} \text{Baum-Welch} \\ \text{New idea: Use of TL information} \end{cases} \end{cases}$$

$\boxed{tagged \text{ corpus}} \rightarrow$ supervised

$\boxed{untagged \text{ corpus}} \rightarrow$ unsupervised

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)

$$
\text{SL text} \longrightarrow \boxed{\text{Segmentation}} \longrightarrow
\begin{array}{l}
\text{segment } s_1 \\
\text{segment } s_2 \\
\quad\quad \vdots \\
\text{segment } s_n
\end{array}
$$

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

SL text $\longrightarrow$ $\boxed{\text{Segmentation}}$ $\rightarrow$

segment s_1
segment s_2
$\vdots$
segment s_n

seg. s_i $\rightarrow$

disambiguations

path g_1
path g_2
$\vdots$
path g_m

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

SL text $\longrightarrow$ $\boxed{\text{Segmentation}}$ $\longrightarrow$ segment s_1, segment s_2, $\vdots$, segment s_n

seg. s_i $\longrightarrow$

disambiguations

path g_1

path g_2

$\vdots$

path g_m

$\longrightarrow$ $\boxed{\text{MT}}$ $\longrightarrow$

translations

$\tau(g_1, s)$

$\tau(g_2, s)$

$\vdots$

$\tau(g_m, s)$

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

SL text $\longrightarrow$ $\boxed{\text{Segmentation}}$ $\nearrow \vdots \searrow$

segment s_1
segment s_2
$\vdots$
segment s_n

seg. s_i $\nearrow \vdots \searrow$

disambiguations

path g_1
path g_2
$\vdots$
path g_m

$\searrow \vdots \nearrow$ $\boxed{\text{MT}}$ $\nearrow \vdots \searrow$

translations

$\tau(g_1, s)$
$\tau(g_2, s)$
$\vdots$
$\tau(g_m, s)$

$\searrow \vdots \nearrow$ $\boxed{\text{TL model}}$ $\nearrow \vdots \searrow$

likelihoods

$p_{\mathsf{TL}}(\tau(g_1, s))$
$p_{\mathsf{TL}}(\tau(g_2, s))$
$\vdots$
$p_{\mathsf{TL}}(\tau(g_m, s))$

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

SL text $\longrightarrow$ $\boxed{\text{Segmentation}}$ $\rightarrow$ segment s_1, segment s_2, $\vdots$, segment s_n

	disambiguations		**translations**		**likelihoods**	
seg. s_i $\rightarrow$	path g_1	$\rightarrow$ $\boxed{\text{MT}}$ $\rightarrow$	$\tau(g_1, s)$	$\rightarrow$ $\boxed{\text{TL model}}$ $\rightarrow$	$p_{\mathsf{TL}}(\tau(g_1, s))$	$\dashrightarrow$ $p(g_1\|s)$
	path g_2		$\tau(g_2, s)$		$p_{\mathsf{TL}}(\tau(g_2, s))$	$\dashrightarrow$ $p(g_2\|s)$
	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	path g_m		$\tau(g_m, s)$		$p_{\mathsf{TL}}(\tau(g_m, s))$	$\dashrightarrow$ $p(g_m\|s)$

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

$s \equiv$	y	la	$para$	si	
	$\{$ CNJ $\}$	$\left\{\begin{array}{c}\text{ART}\\ \text{PRN}\end{array}\right\}$	$\left\{\begin{array}{c}\text{VB}\\ \text{PR}\end{array}\right\}$	$\{$ CNJ $\}$	
					$p(g_i\|s)$
$g_1 \equiv$	CNJ	ART	PR	CNJ	
$\tau(g_1, s) \equiv$	i (and)	la (the)	$per\ a$ (for/to)	si (if)	0.0001
$g_2 \equiv$	CNJ	ART	VB	CNJ	
$\tau(g_2, s) \equiv$	i (and)	la (the)	$para$ (stop)	si (if)	0.4999
$g_3 \equiv$	CNJ	PRN	PR	CNJ	
$\tau(g_3, s) \equiv$	i (and)	la (it/her)	$per\ a$ (for/to)	si (if)	0.0001
$g_4 \equiv$	CNJ	PRN	VB	CNJ	
$\tau(g_4, s) \equiv$	i (and)	la (it/her)	$para$ (stop)	si (if)	0.4999

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

$s \equiv$	y	la	$para$	si	
	{ CNJ }	{ ART, PRN }	{ VB, PR }	{ CNJ }	$p(g_i\|s)$
$g_1 \equiv$	CNJ	ART	PR	CNJ	
$\tau(g_1, s) \equiv$	i (and)	la (the)	$per\ a$ (for/to)	si (if)	0.0001
$g_2 \equiv$	CNJ	ART	VB	CNJ	
$\tau(g_2, s) \equiv$	i (and)	la (the)	$para$ (stop)	si (if)	0.4999
$g_3 \equiv$	CNJ	PRN	PR	CNJ	
$\tau(g_3, s) \equiv$	i (and)	la (it/her)	$per\ a$ (for/to)	si (if)	0.0001
$g_4 \equiv$	CNJ	PRN	VB	CNJ	
$\tau(g_4, s) \equiv$	i (and)	la (it/her)	$para$ (stop)	si (if)	0.4999

Free ride: word translated the same way independently of the tag selected

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

$$p(g_i|s) \propto p(g_i|\tau(g_i,s))\, p_{\mathsf{TL}}(\tau(g_i,s))$$

- $p(g_i|s)$: Probability of path g_i to be the correct disambiguation of segment s
- $p_{\mathsf{TL}}(\tau(g_i,s))$: 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 prevoius idea ...

Cooperative learning of HMM (I)

- Use of the prevoius idea ...
- Bidirectional MT system translating between languages A and B

Cooperative learning of HMM (I)

- Use of the prevoius 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 prevoius 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 prevoius 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)

Lang. A	Lang. B

$$
\begin{tikzcd}
 & M_B[0] \arrow[dl] \\
M_A[1] \arrow[r] & M_B[1] \arrow[dl] \\
M_A[2] \arrow[r] & M_B[2] \\
\vdots & \vdots \\
 & M_B[k-1] \arrow[dl] \\
M_A[k] \arrow[r] & M_B[k] \\
\vdots & \vdots
\end{tikzcd}
$$

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

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%)

	Avg. PoS error		**Avg. translation error**		**Avg. It.**
	Spanish	Catalan	Spanish	Catalan	
"good start" →	24.9%	27.5%	6.2%	6.7%	2
"bad start" →	25.9%	26.4%	6.1%	6.8%	5
Baum-Welch →	31.7%	37.8%	8.4%	13.6%	14
supervised →	10.4%	16.5%	2.6%	3.0%	

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 compare 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 compare 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 !!