Exploring the use of target-language information to train the part-of-speech tagger of machine translation systems*

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: determining the lexical category or PoS of each word that appears in a text

Lexically ambiguous word: word with more than one possible lexical category or part-of-speech (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 intermediate representation (IR), transformations are applied and, finally, target language (TL) text is generated

$$\text{SL text} \longrightarrow \boxed{\text{Analysis}} \xrightarrow{\text{SLIR}} \boxed{\text{Transformation}} \xrightarrow{\text{TLIR}} \boxed{\text{Generation}} \longrightarrow \text{TL text}$$

- 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
la calle	*la* (article)	*el carrer* (the street)
	la (pronoun)	* *la carrer* (it/her street)

←gender agreement rule applied

PoS tagging with HMM (I)

Classical 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 (set of all possible PoS tags for a given word)

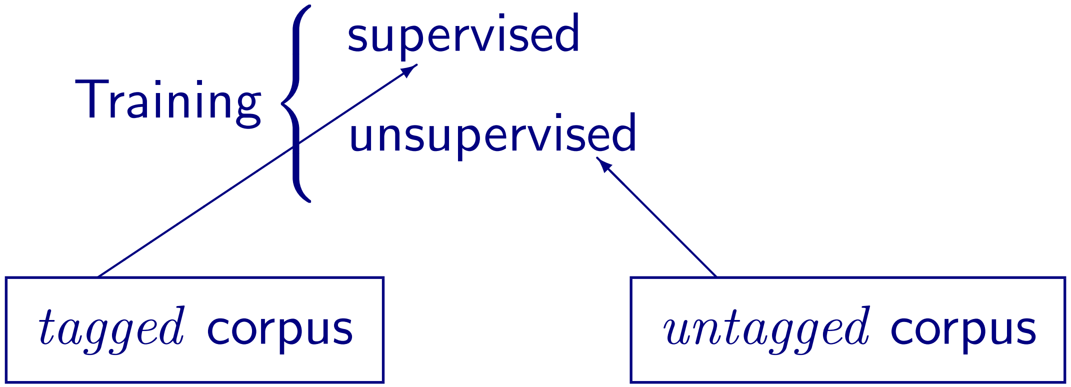
PoS tagging with HMM (II)

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:

Training
- supervised
- unsupervised
  - Baum-Welch
  - New idea: Use of TL information

tagged corpus

untagged corpus

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

Training as if we had a tagged corpus:

- Transition probabilities

$$a_{\gamma_i\gamma_j} = \frac{\tilde{n}(\gamma_i\gamma_j)}{\sum_{\gamma_k\in\Gamma}\tilde{n}(\gamma_i\gamma_k)}, \text{ where } \gamma_i \text{ is a tag}$$

- Emission probabilities

$$b_{\gamma_i\sigma} = \frac{\tilde{n}(\sigma,\gamma_i)}{\sum_{\sigma':\gamma_i\in\sigma'}\tilde{n}(\sigma',\gamma_i)}, \text{ where } \sigma \text{ is an ambiguity class}$$

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

SL text

$\downarrow$

$\boxed{\text{segmentation}}$

$\downarrow$

seg. s_1, seg. s_2, seg. s_3 . . . seg. s_n

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

SL text

$\downarrow$

$\boxed{\text{segmentation}}$

$\downarrow$

seg. s_1, seg. s_2, seg. s_3 $\ldots$ seg. s_n

disambiguations

$$\text{seg. } s_i \begin{array}{l} \nearrow \\ \vdots \\ \searrow \end{array} \begin{array}{l} \text{path } g_1 \\ \text{path } g_2 \\ \vdots \\ \text{path } g_m \end{array}$$

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

SL text

↓

$\boxed{\text{segmentation}}$

↓

seg. s_1, seg. s_2, seg. s_3 ... seg. s_n

	disambiguations				**translations**
	path g_1				$\tau(g_1, s)$
seg. s_i	path g_2		$\boxed{\text{MT}}$		$\tau(g_2, s)$
	⋮				⋮
	path g_m				$\tau(g_m, s)$

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

SL text

↓

$\boxed{\text{segmentation}}$

↓

seg. s_1, seg. s_2, seg. s_3 ... seg. s_n

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

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

SL text

↓

segmentation

↓

seg. s_1, seg. s_2, seg. s_3 . . . seg. s_n

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

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

$$s \equiv \quad y \quad la \quad para \quad si$$

$$\{\ \text{CNJ}\ \} \quad \left\{ \begin{array}{c} \text{ART} \\ \text{PRN} \end{array} \right\} \quad \left\{ \begin{array}{c} \text{VB} \\ \text{PR} \end{array} \right\} \quad \{\ \text{CNJ}\ \}$$

					$p(g_i\|s)$
$g_1 \equiv$	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 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
  - ...
- $p(g_i|\tau(g_i,s))$: Contribution of the disambiguation path g_i to the translation given by $\tau(g_i,s)$

Target-language model

- Trigram model of TL surface forms (words as they appear in raw text)
- Probabilities smoothed via *deleted interpolation* and Good-Turing
- Likelihood evaluation of a segment:
  - taking into account the two preceding words of the segment, and
  - taking into account the two first words of the next segment
- Problem: Shorter translations receive higher scores than larger ones

Experiments (I)

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

- Translating from Spanish to Catalan

- Catalan trigram language model from $1\,822\,067$-word corpus

- Use of three different corpora with $200\,000$ words each

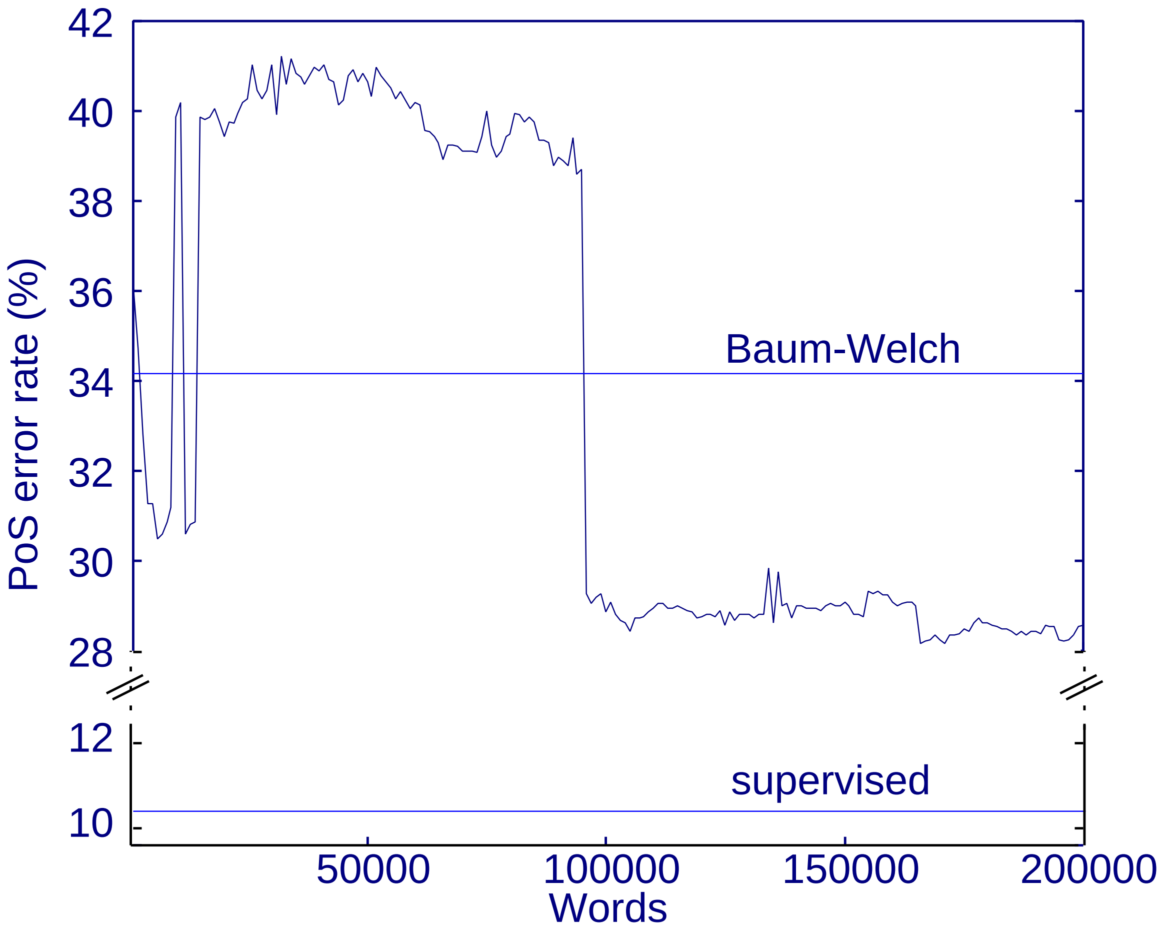
- We calculate the HMM parameters after every $1\,000$ words

Experiments (II)

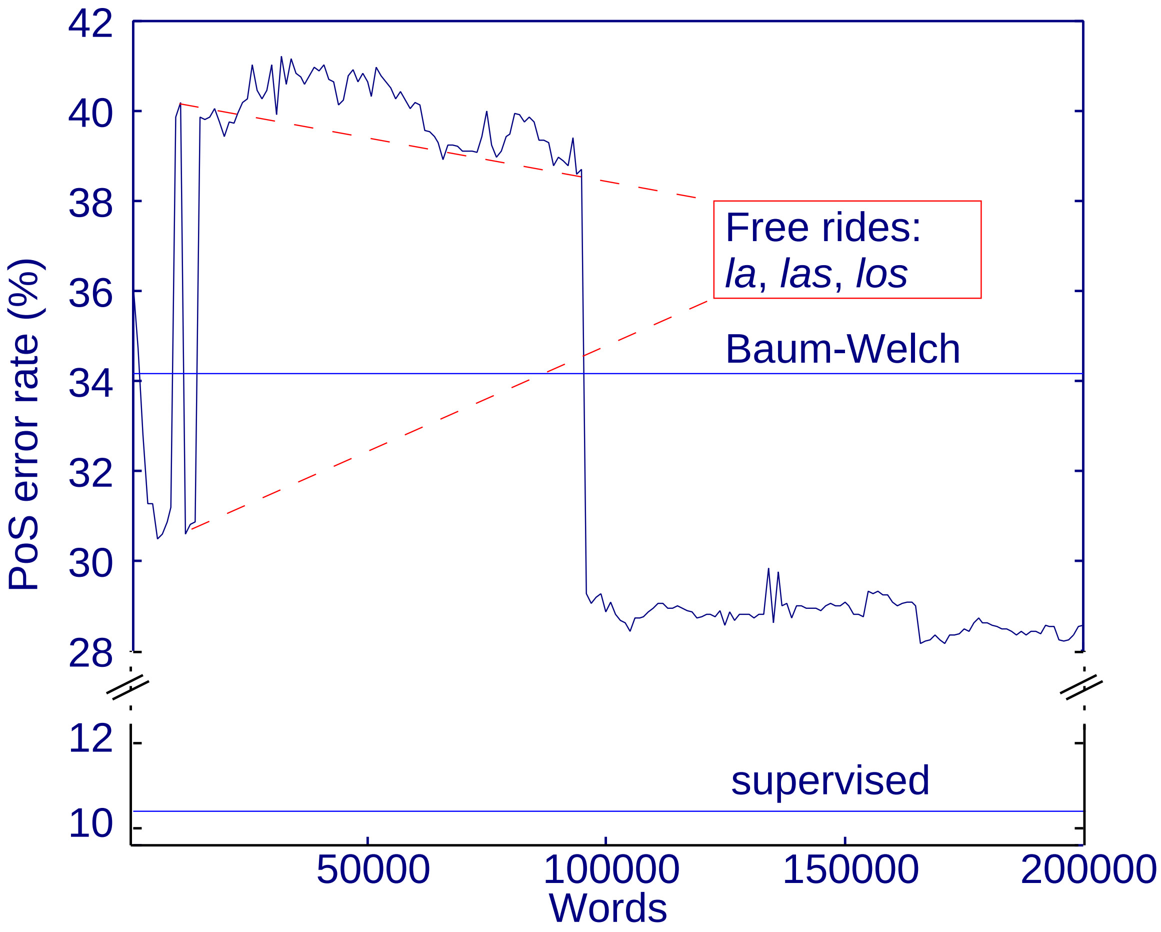
- Performance measures with an independent Spanish corpus:
  - PoS error rate with $8\,031$-word hand-tagged corpus
  - Translation error rate with human corrected translations

- For comparison purpose:
  - HMM-based PoS tagger trained from $1\,000\,000$-word Spanish untagged corpus with the Baum-Welch algorithm (unsupervised)
  - HMM-based PoS tagger trained from $20\,000$-word Spanish hand-tagged corpus (supervised)

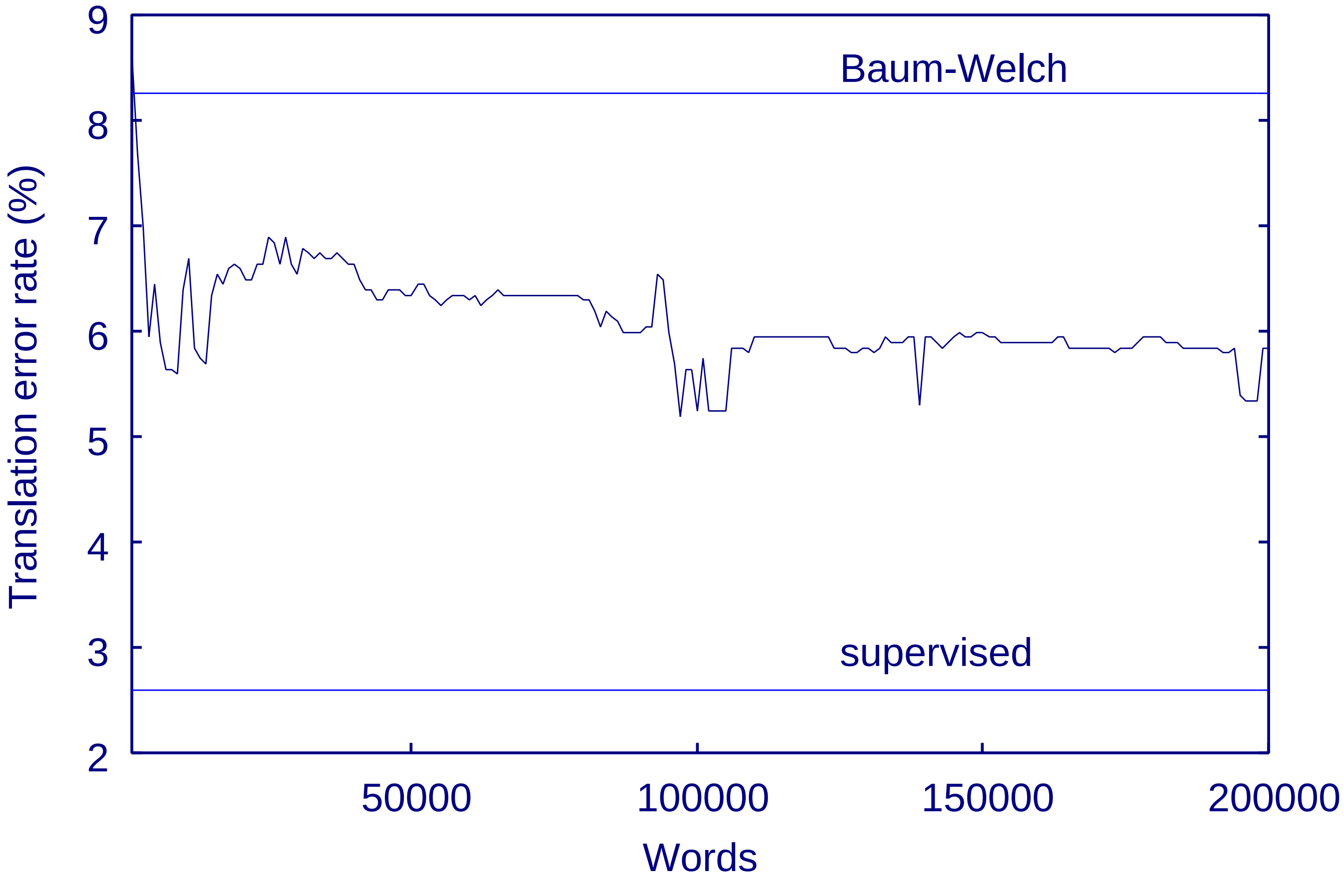
Results: PoS error



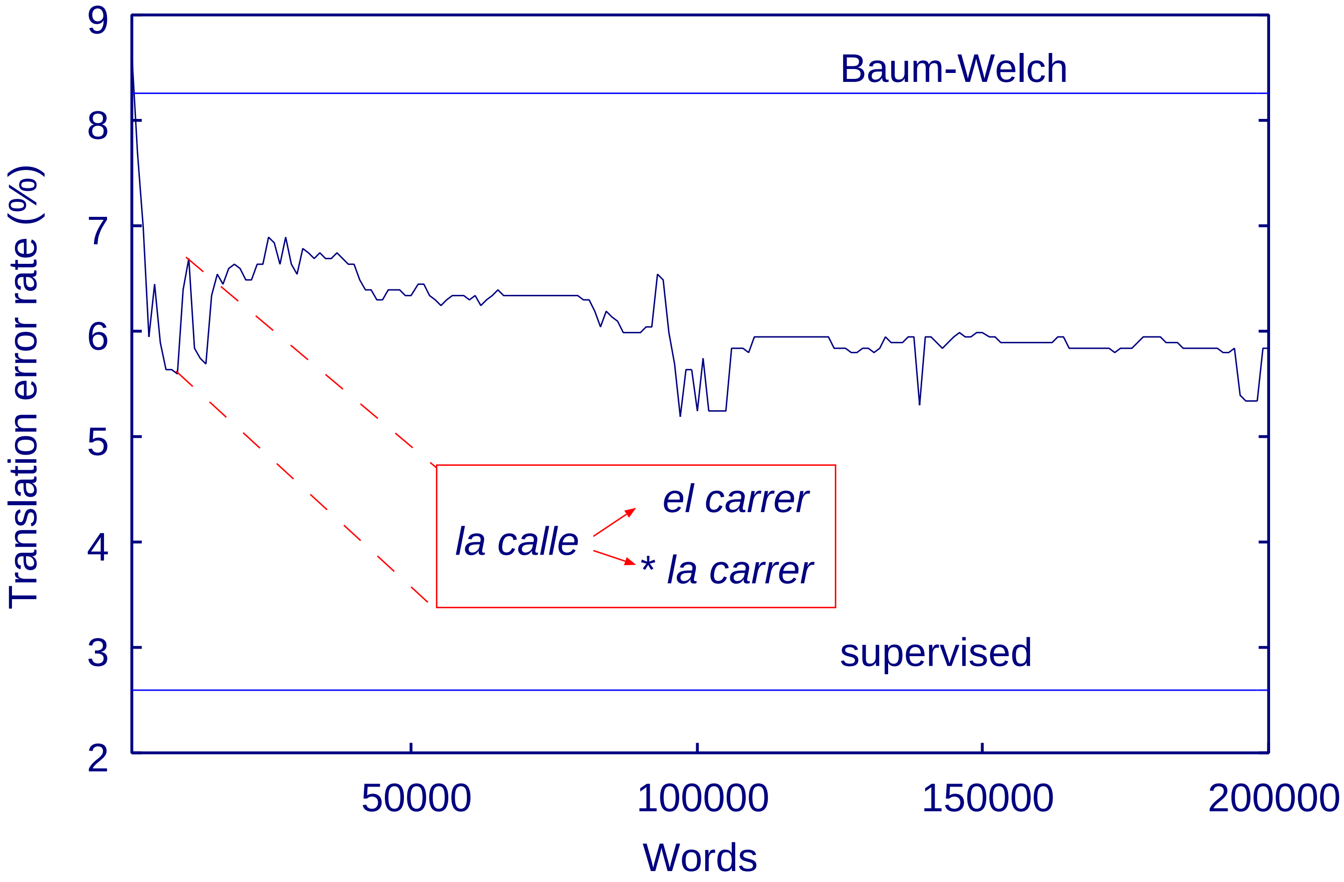
Results: PoS error



Results: Translation error



Results: Translation error



Results: Reducing the impact of free rides

Common free rides: *la*, *las*, *los*

- 6.14% of all words, and 22.98% of ambiguous words
- Ambiguity class: ART
PRN

Results: Reducing the impact of free rides

Common free rides: la, las, los

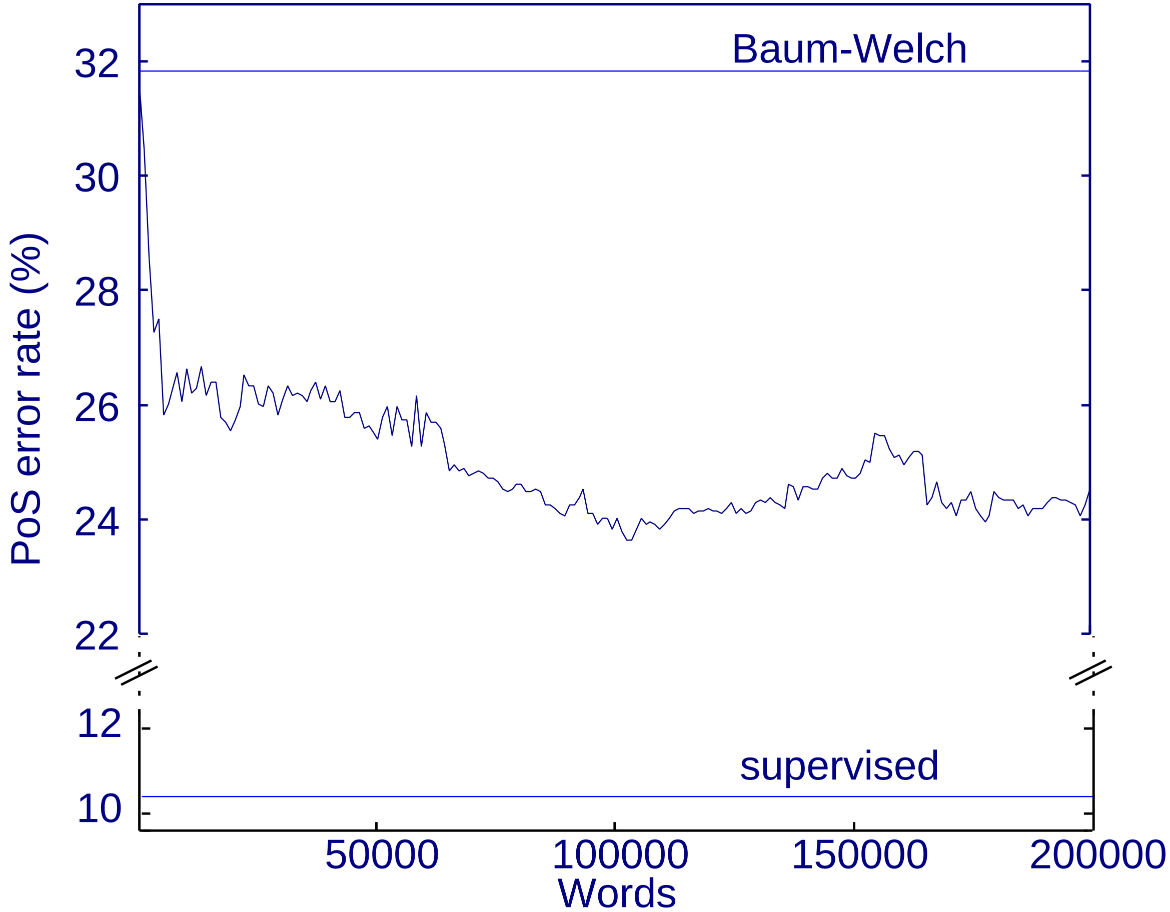
- 6.14% of all words, and 22.98% of ambiguous words
- Ambiguity class: ART
PRN

Solution: Use of linguistic information. Some impossible tag bigrams are forbidden

- We forbid, for example:
  - article or preposition before verb in personal form
  - article before proclitic pronouns

Use: Do not take into account disambiguation paths with one or more forbidden bigram

Results: Reducing the impact of free rides (PoS error)



Discussion

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

Discussion

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

- The presence of free rides make the algorithm behaves unstably due to the kind of TL model used
  - The problem is partially solved using an small amount of linguistic information

Discussion

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

- The presence of free rides make the algorithm behaves unstably due to the kind of TL model used
  - The problem is partially solved using an small amount of linguistic information

- Reduction of the translation error rate around 2% with an small amount of text, even when no linguistic information was used

Discussion

- PoS error and translation error rates lie between those produce by supervised and unsupervised methods
- The presence of free rides make the algorithm behaves unstably due to the kind of TL model used
  - The problem is partially solved using an small amount of linguistic information
- Reduction of the translation error rate around 2% with an small amount of text, even when no linguistic information was used
- The training method produces PoS tagger that are 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