Speeding up target-language driven part-of-speech tagger training for machine translation

Felipe Sánchez-Martínez Juan Antonio Pérez-Ortiz
Mikel L. Forcada

Transducens Group – Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain

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

MICAI, 5th Mexican International Conference on Artificial Intelligence
Apizaco, México
November 16, 2006

Outline

Outline

Part-of-speech tagging

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

- Lexically ambiguous word: word with more than one possible lexical category or PoS

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

- Ambiguities are usually solved according to the surrounding context

PoS tagging for machine translation /1

Indirect rule-base machine translation (MT) systems usually perform PoS tagging as a subtask of the analysis procedure

source text → Analysis → Transfer → Generation → target text

PoS tagging for machine translation /2

PoS tagging becomes crucial

- Translation may differ from one PoS to another

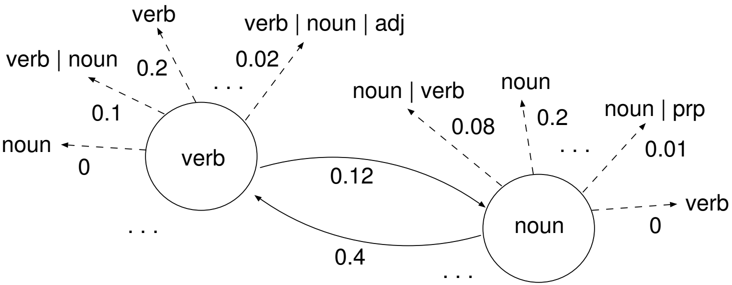
English	**PoS**	**Spanish**
book	noun	*libro*
	verb	*reservar*

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

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

PoS tagging with HMM

- *Hidden Markov models* are one of the standard statistical solutions for PoS tagging



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

HMM parameter estimation

- *Supervisedly* (non-ambiguous corpora available):
  - Maximum-likelihood estimate (MLE)

- *Unsupervisedly* (only ambiguous corpora available):
  - Baum-Welch (Expectation-maximization, EM)
  - Our recently proposed (Sánchez-Martínez et al. 2004) *target-language (TL) driven* method
  - ...

Outline

Target-language driven method overview

- The method uses the MT system in which the resulting tagger will be embedded; however it will also work for other natural language processing tasks
- A target-language (TL) model is used to choose the best disambiguations
- HMM parameters are calculated according to the likelihood of the corresponding translations into TL
- The resulting tagger is tuned to the translation quality

Example

- Source-language (SL) sentence (English):
  - He-prn books-noun|verb the-art room-noun|verb

- Possible translations (Spanish) according to each disambiguation and their normalized likelihoods according to a target-language (TL) model:
  - Él-prn reserva-verb la-art habitación-noun 0.75
  - Él-prn reserva-verb la-art aloja-verb 0.15
  - Él-prn libros-noun la-art habitación-noun 0.06
  - Él-prn libros-noun la-art aloja-verb + 0.04

  1.00

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

Example

- Source-language (SL) sentence (English):
  - He-prn books-noun|verb the-art room-noun|verb

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

• Él-prn reserva-verb la-art habitación-noun	0.75
• Él-prn reserva-verb la-art aloja-verb	0.15
• Él-prn libros-noun la-art habitación-noun	0.06
• Él-prn libros-noun la-art aloja-verb	+ 0.04
	1.00

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

Example

- Source-language (SL) sentence (English):
  - He-prn books-noun|verb the-art room-noun|verb

- Possible translations (Spanish) according to each disambiguation and their normalized likelihoods according to a target-language (TL) model:
  - Él-prn reserva-verb la-art habitación-noun 0.75
  - Él-prn reserva-verb la-art aloja-verb 0.15
  - Él-prn libros-noun la-art habitación-noun 0.06
  - Él-prn libros-noun la-art aloja-verb + 0.04

  1.00

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

Disadvantage

- The number of possible disambiguations to translate grows exponentially with the segment length
- Translation is the most time-consuming task
- Consequence: Segment length must be constrained to keep complexity under control
  - Potential benefits of likelihood estimated from longer segments is rejected
- Goal: To overcome this problem
- How? Pruning unlikely disambiguation paths by using *a priori* knowledge

Outline

Pruning method /1

- Based on an initial model of SL tags (M_{tag})
- Assumption: Any reasonable model of SL tags may be useful to choose a set of possible disambiguation paths, being the correct one in that set
  - It is not necessary to translate all possible disambiguation paths, but the “promising” ones
- The model used for pruning can be update dynamically

Pruning method /2

1. The *a priori* likelihood $p(\boldsymbol{g_i}|s, M_{\mathrm{tag}})$ of each possible disambiguation path $\boldsymbol{g_i}$ of segment s is calculated using the model M_{tag}
2. Then, the set of disambiguation paths to take into account is determined:
   - Only the most likely disambiguation paths
   - A mass probability threshold ρ is introduced
   - The set of disambiguation paths taken into account satisfies

$$\rho \leq \sum_{\forall \boldsymbol{g_i} \in T(s)} p(\boldsymbol{g_i}|s, M_{\mathrm{tag}})$$

HMM updating

- The model M_{tag} used for pruning can be updated with the new evidences collected from the TL
- The update consist of:
  1. Calculating the HMM parameters with the counts collected from the TL
  2. Mixing the parameters of the new HMM with the initial one

HMM parameters mixing

- Let $\boldsymbol{\theta} = (a_{\gamma_1\gamma_1}, ..., a_{\gamma_{|\Gamma|}\gamma_{|\Gamma|}}, b_{\gamma_1\sigma_1}, ..., b_{\gamma_{|\Gamma|}\sigma_{|\Sigma|}})$ be a vector containing all the parameters of a given HMM
- Mixing equation:

$$\boldsymbol{\theta}_{\textbf{mixed}}(x) = \lambda(x)\,\boldsymbol{\theta}_{\textbf{TL}}(x) + (1 - \lambda(x))\,\boldsymbol{\theta}_{\textbf{init}}$$

- $\lambda(x)$ assigns a weight to the model estimated using the counts collected from the TL ($\boldsymbol{\theta}_{\textbf{TL}}$)
  - This weight function is made to depend on the number x of SL words processed so far

$$\lambda(x) = x/C$$

Outline

Overview

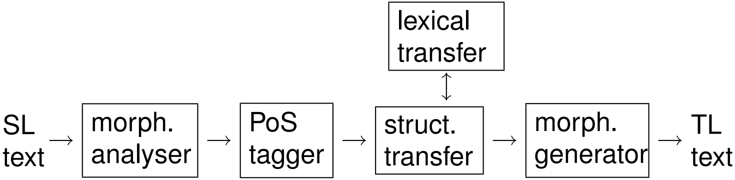
- Task: Training a Spanish PoS tagger Catalan being the TL
- TL model: Trigram language model trained from a Catalan corpus with around 2 000 000 words
- SL corpora: 5 Spanish disjoint corpora of 500 000 words
- Initial model: estimated through Kupiec's method
- HMM updating: after every 1 000 words
- Mass probability threshold: $0.1 \leq \rho \leq 1.0$, increment: 0.1
- Evaluation: hand-tagged corpus with around 8 000 words

Framework

- Open-source shallow transfer MT engine Apertium, `http://apertium.org`
- Packages: `lttoolbox-1.0.1`, `apertium-1.0.1`, `apertium-es-ca-1.0.1`
- The method presented (including the language model) is implemented inside package `apertium-tagger-training-tools`
- All packages, including source code, can be freely downloaded from `http://sourceforge.net/projects/apertium`

Apertium working scheme

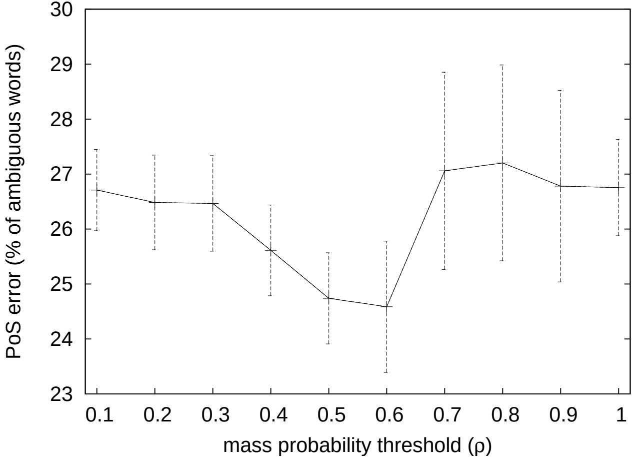
- Shallow-transfer machine translation architecture



- PoS tagger is trained by using the rest of the modules of the MT engine after it
- The morphological analyzer is used to preprocess SL texts

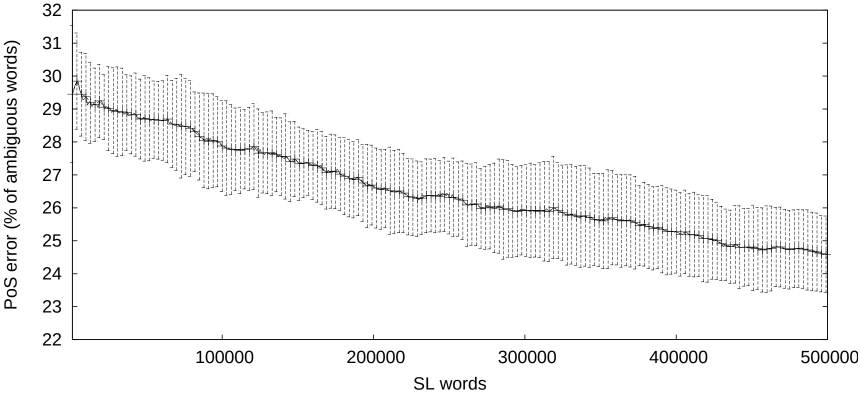
Results /1

Mean and std. dev. of the PoS tagging error rate achieved after training for each value of ρ



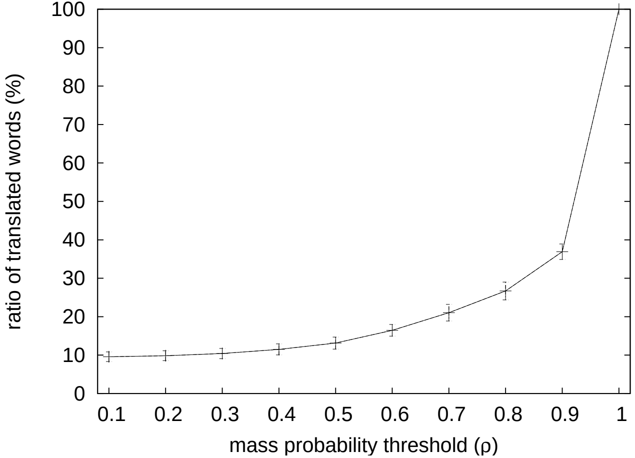
Results /2

Evolution of the mean and std. dev. of the PoS tagging error rate of the mixed model used for pruning for $\rho = 0.6$



Results /3

Percentage of translated words for each value of ρ



Outline

Concluding remarks

- The pruning method avoids more than 80% of the translations to perform
- The results achieved are even better than when no pruning is performed, when $\rho = 1.0$
    - HMM parameters involved in those discarded disambiguations have a null count
    - When no pruning is done their counts are small but never null

Future work

- Try other weighting functions giving earlier a higher weight to the model being learned from the TL
  - Test how fast the TL-driven method learns
- Test two additional strategies to select the disambiguation paths to take into account
  - Dynamically change the value of the mass probability threshold ρ while training
  - Instead of using ρ, always select a fix number *k* of disambiguation paths to translate

Further reading

- Sánchez-Martínez F., J.A. Pérez-Ortiz and M. L. Forcada
  Exploring the use of target-language information to train the part-of-speech tagger of machine translation systems
  Lecture Notes in Computer Science 3230 (Advances in Natural Language Processing, Proceedings of EsTAL - España for Natural Language Processing), p. 137–148, 2004.
- Corbí-Bellot A. M., M. L. Forcada, S. Ortiz-Rojas, J. A. Pérez-Ortiz, G. Ramírez-Sánchez, F. Sánchez-Martínez, I. Alegria, A. Mayor, K. Sarasola
  An open-source shallow-transfer machine translation engine for the Romance languages of Spain
  Proceedings of the Tenth Conference of the European Associtation for Machine Translation, p. 79–80, 2005.

Speeding up target-language driven part-of-speech tagger training for machine translation

Felipe Sánchez-Martínez   Juan Antonio Pérez-Ortiz
Mikel L. Forcada

Transducens Group – Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain

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

MICAI, 5th Mexican International Conference on Artificial Intelligence
Apizaco, México
November 16, 2006