Training Part-of-Speech Taggers to build Machine Translation Systems for Less-Resourced Language Pairs

Felipe Sánchez-Martínez, Carme Armentano-Oller, Juan Antonio Pérez-Ortiz, Mikel L. Forcada

Transducens Group
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant
E-03071 Alacant, Spain
{fsanchez,carmentano,japerez,mlf}@dlsi.ua.es

Resumen: Este articulo revisa el empleo de un método no supervisado para la obtención de desambiguadores léxicos categoriales para su empleo dentro del ingenio de traducción automática (TA) de código abierto Apertium. El método emplea el resto de módulos del sistema de TA y un modelo de la lengua destino de la traducción para la obtención de desambiguadores léxicos categoriales que después se usan dentro de la plataforma de TA Apertium para traducir. Los experimentos realizados con el par de lenguas occitano–catalán (un caso de estudio para pares de lenguas minorizadas con pocos recursos) muestran que la cantidad de corpus necesario para el entrenamiento es reducida comparado con los tamaños de corpus habitualmente usados con otros métodos de entrenamiento no supervisado como el algoritmo de Baum y Welch. Esto hace que el método sea especialmente apropiado para la obtención de desambiguadores léxicos categoriales para su empleo en TA entre pares de lenguas minorizadas. Además, la calidad de traducción del sistema de TA que utiliza el desambiguador léxico categorial resultante es comparativamente mejor.
Palabras clave: traducción automática, lenguas minorizadas, desambiguación léxica categorial, modelos ocultos de Markov

Abstract: In this paper we review an unsupervised method that can be used to train the hidden-Markov-model-based part-of-speech taggers used within the open-source shallow-transfer machine translation (MT) engine Apertium. This method uses the remaining modules of the MT engine and a target language model to obtain part-of-speech taggers that are then used within the Apertium MT engine in order to produce translations. The experimental results on the Occitan–Catalan language pair (a case study of a less-resourced language pair) show that the amount of corpora needed by this training method is small compared with the usual corpus sizes needed by the standard (unsupervised) Baum-Welch algorithm. This makes the method appropriate to train part-of-speech taggers to be used in MT for less-resourced language pairs. Moreover, the translation performance of the MT system embedding the resulting part-of-speech tagger is comparatively better.
Keywords: machine translation, less-resourced languages, part-of-speech tagging, hidden Markov models

1 Introduction

The growing availability of machine-readable (monolingual and parallel) corpora has given rise to the development of real applications such as corpus-based machine translation (MT). However, when MT involves less-resourced language pairs, such as Occitan–Catalan (see below), the amount of monolingual or parallel corpora, if available, is not enough to build a general-purpose open-domain MT system (Forcada, 2006). In these cases the only realistic approach to attain high performance in general translation is to follow a rule-based approach, but at the expense of the large costs needed for building the necessary linguistic resources (Arnold, 2003).

In this paper we focus on the training of the hidden Markov model (HMM)-based part-of-speech taggers used by a particular open-source Occitan–Catalan MT system (Armentano-Oller and Forcada, 2006), that has been built using Apertium, an open-source platform for building MT systems (see section 2). Occitan–Catalan is an interesting example of a less-resourced language pair. HMMs are a common statistical approach to part-of-speech tagging, but they usually demand large corpora, which are seldom available for less-resourced languages.

Catalan is a Romance language spoken by around 6 million people, mainly in Spain (where it is co-official in some regions), but also in Andorra (where it is the official language), in parts of Southern France and in the Sardinian city of l'Alguer (Alghero).

Occitan, also known as *lenga d'òc* or *langue d'oc*, is also a Romance language, but with a reduced community of native speakers. It is reported to have about one million speakers, mainly in Southern France, but also in some valleys of Italy and in the Val d'Aran, a small valley of the Pyrenees of Catalonia, inside the territory of Spain. This last variety is called Aranese; all of the experiments reported here have been performed with the Aranese variety of Occitan.

Although Occitan was one of the main literary languages in Medieval Europe, nowadays it is legally recognized only in the Val d'Aran, where it has a limited status of cooficiality. In addition, Occitan dialects have strong differences, and its standardization as a single language still faces a number of open issues. Furthermore, the lack of general-purpose machine-readable texts restricts the design and construction of natural-language processing applications such as part-of-speech taggers. The Apertium-based Occitan–Catalan MT system (Armentano-Oller and Forcada, 2006) mentioned along this paper has been built to translate into the Occitan variety spoken in the Val d'Aran, called Aranese, which is a sub-dialect of Gascon (one of the main dialects of Occitan).

When part-of-speech tagging is viewed as an intermediate task for the translation process the use in a unsupervised manner of target-language (TL) information, in addition to the source language (SL), has been shown to give better results than the standard (also unsupervised) Baum-Welch algorithm (Sánchez-Martínez, Pérez-Ortiz, and Forcada, 2004b). Moreover, as the experimental results show, the amount of source language text is small compared with corpus sizes needed by the standard Baum-Welch algorithm. Because of this, it may be said that this training method is specially suited to train part-of-speech taggers to be embedded in MT systems involving less-resourced language pairs.

Carbonell et al. (2006) proposed a new MT framework in which a large full-form bilingual dictionary and a huge TL corpus is used to carry out the translation; neither parallel corpora nor transfer rules are needed. The idea behind Carbonell's paper and that of the method we present here share the same principle: if the goal is to get good translations into TL, let TL decides whether a given "construction" in the TL is good or not. In contrast, Carbonell's method uses TL information at translation time, while ours uses only TL information when training one module that is then used to carry out the translation; therefore, no TL information is used by our method at translation time.

The rest of the paper is organized as follows: section 2 overviews the open-source platform for building MT systems Apertium; next, in section 3 the TL-driven training method used to train the Occitan part-of-speech tagger is introduced; section 4 shows the experiments and the results achieved; finally in section 5 we discuss the method and the results achieved.

2 Overview of Apertium

Apertium[1] (Armentano-Oller et al., 2006; Corbí-Bellot et al., 2005) is an open-source platform for developing MT systems, initially intended for related language pairs. The Apertium MT engine follows a shallow transfer approach and may be seen as an assembly line consisting of the following modules (see figure 1):

- A *de-formatter* which separates the text to be translated from the format information (RTF and HTML tags, whitespace, etc.). Format information is encapsulated so that the rest of the modules treat it as blanks between words.

[1]The MT engine, documentation, and linguistic data for different language pairs can be downloaded from `http://apertium.sf.net`.

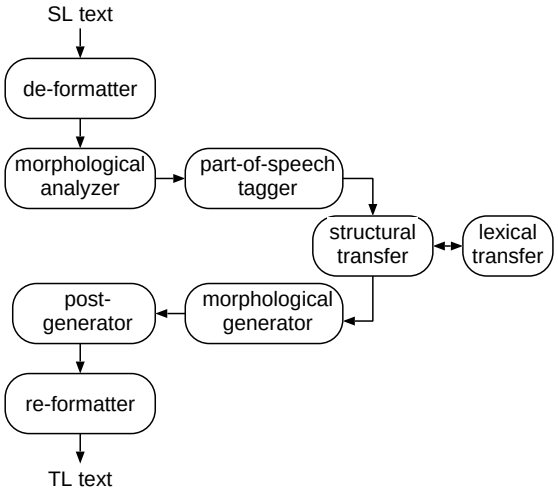

Figure 1: Modules of the Apertium shallow-transfer MT platform (see section 2).

- A *morphological analyzer* which tokenizes the SL text in surface forms and delivers, for each surface form, one or more *lexical forms* consisting of *lemma*, *lexical category* and morphological inflection information.
- A *part-of-speech tagger* which chooses, using a first-order hidden Markov model (HMM) (Cutting et al., 1992), one of the lexical forms corresponding to an ambiguous surface form. This is the module whose training is discussed in section 3.
- A *lexical transfer* module which reads each SL lexical form and delivers the corresponding TL lexical form by looking it up in a bilingual dictionary.
- A *structural shallow transfer* module (parallel to the lexical transfer) which uses a finite-state chunker to detect patterns of lexical forms which need to be processed for word reorderings, agreement, etc., and then performs these operations.[2]
- A *morphological generator* which delivers a TL surface form for each TL lexical form, by suitably inflecting it.
- A *post-generator* which performs orthographic operations such as contractions (e.g. Spanish *del=de+el*) and apostrophations (e.g. Catalan *l'institut=el+institut*).
- A *re-formatter* which restores the format information encapsulated by the de-formatter into the translated text.

Modules use text to communicate, which makes it much easier to diagnose or modify the behavior of the system.

2.1 Linguistic data and compilers

The Apertium MT engine is completely independent from the linguistic data used for translating between a particular pair of languages.

Linguistic data is coded using XML-based formats;[3] this allows for interoperability, and for easy data transformation and maintenance. In particular, files coding linguistic data can be automatically generated by third-party tools.

Apertium provides compilers to convert the linguistic data into the corresponding efficient form used by each module of the engine. Two main compilers are used: one for the four lexical processing modules (morphological analyzer, lexical transfer, morphological generator, and post-generator) and another one for the structural transfer. The first one generates finite-state letter transducers (Garrido-Alenda, Forcada, and Carrasco, 2002) which efficiently code the lexical data; the last one uses finite-state machines to speed up pattern matching. The use of such efficient compiled data formats makes the engine capable of translating tens of thousands of words per second in a current desktop computer.

3 Target-language-driven part-of-speech tagger training

This section overviews the TL-driven training method that has been used to unsupervisedly train the HMM-based Occitan part-of-speech tagger used within the Apertium-based Occitan–Catalan MT system (Armentano-Oller et al., 2006). For a deeper description we refer the reader to papers by Sánchez-Martínez et al. (Sánchez-Martínez, Pérez-Ortiz, and Forcada, 2004b;

[2]This describes Apertium Level 1, used for the experiments in this paper; in Apertium Level 2, currently being used for less-related pairs, a three-stage structural transfer is used to perform inter-chunk operations.

[3]The XML formats (`http://www.w3.org/XML/`) for each type of linguistic data are defined through conveniently-designed XML document-type definitions (DTDs) which may be found inside the `apertium` package.

Sánchez-Martínez, Pérez-Ortiz, and Forcada, 2004a; Sánchez-Martínez, Pérez-Ortiz, and Forcada, 2006).

Typically, the training of general purpose HMM-based part-of-speech taggers is done using the *maximum-likelihood estimate* (MLE) method (Gale and Church, 1990) when tagged corpora[4] are available (supervised method), or using the Baum-Welch algorithm (Cutting et al., 1992; Baum, 1972) with untagged corpora[5] (unsupervised method). However, if the part-of-speech tagger is to be embedded as a module in a MT system, as is the case, HMM training can be done in an unsupervised manner by using some modules of the MT system and information from both SL and TL.

The main idea behind the use of TL information is that the correct disambiguation (tag assignment) of a given SL segment will produce a more likely TL translation than any (or most) of the remaining wrong disambiguations. In order to apply this method these steps are followed:

- first the SL text is split into adequate segments (so that they are small and independently translated by the rest of the MT engine); then,
- all possible disambiguations for each text segment are generated and translated into the TL; after that,
- a statistical TL model is used to compute the likelihood of the translation of each disambiguation; and,
- these likelihoods are used to adjust the parameters of the SL HMM: the higher the likelihood, the higher the probability of the original SL tag sequence in the HMM being trained.

The way this training method works can be illustrated with the following example. Suppose that we are training an English PoS tagger to be used within a rule-based MT system translating from English to Spanish, and that we have the following segment in English, $s =$ *"He books the room"*. The first step is to use a morphological analyzer to obtain the set of all possible part-of-speech tags for each word. Suppose that the morphological analysis of the previous segment according to the lexicon is: *He* (pronoun), *books* (verb or noun), *the* (article), and *room* (verb or noun). As there are two ambiguous words (*books* and *room*) we have, for the given segment, four disambiguation *paths* or part-of-speech combinations, that is to say:

- $\boldsymbol{g}_1$ = (pronoun, verb, article, noun),
- $\boldsymbol{g}_2$ = (pronoun, verb, article, verb),
- $\boldsymbol{g}_3$ = (pronoun, noun, article, noun), and
- $\boldsymbol{g}_4$ = (pronoun, noun, article, verb).

Let τ be the function representing the translation task. The next step is to translate the SL segment into the TL according to each disambiguation path $\boldsymbol{g}_i$:

- $\tau(\boldsymbol{g}_1, s) =$ "*Él reserva la habitación*",
- $\tau(\boldsymbol{g}_2, s) =$"*Él reserva la aloja*",
- $\tau(\boldsymbol{g}_3, s) =$"*Él libros la habitación*", and
- $\tau(\boldsymbol{g}_4, s) =$"*Él libros la aloja*".

It is expected that a Spanish language model will assign a higher likelihood to translation $\tau(\boldsymbol{g}_1, s)$ than to the other ones, which make little sense in Spanish. As a result, the tag sequence $\boldsymbol{g}_1$ will have a higher probability than the other ones.

To estimate the HMM parameters, the calculated probabilities are used as if fractional counts were available to a supervised training method based on the MLE method in conjunction with a smoothing technique (Sánchez-Martínez, Pérez-Ortiz, and Forcada, 2004b).

As expected, the number of possible disambiguations of a text segment grows exponentially with its length, the translation task being the most time-consuming one. This problem has been successfully addressed (Sánchez-Martínez, Pérez-Ortiz, and Forcada, 2006) by using a very simple pruning method that avoids performing more than 80% of the translations without loss in accuracy.

An implementation of the method described in this section can be downloaded from the Apertium project web page,[6] and

[4] In a *tagged corpus* each occurrence of each word (ambiguous or not) has been assigned the correct part-of-speech tag.

[5] In an *untagged corpus* all words are assigned (using, for instance, a morphological analyzer) the set of all possible part-of-speech tags independently of context without choosing one of them.

[6] `http://apertium.sourceforge.net`. The

may simplify the initial building of Apertium-based MT systems for new language pairs, yielding better tagging results than the Baum-Welch algorithm (Sánchez-Martínez, Pérez-Ortiz, and Forcada, 2004b).

4 Experiments

The method we present is aimed at producing part-of-speech taggers to be used in MT systems. In this section we report the results achieved when training the Occitan part-of-speech tagger of the Apertium-based Occitan–Catalan MT system.[7] Note that when training the Occitan part-of-speech tagger the whole MT engine, except for the part-of-speech tagger itself, is used to produce texts from which statistics about TL (Catalan) will be collected.

Before training, the Occitan corpus is divided into small segments that can be independently translated by the rest of the translation engine. To this end, information about the structural transfer patterns is taken into account. The segmentation is performed at nonambiguous words whose part-of-speech tag is not present in any structural transfer pattern, or at nonambiguous words appearing in patterns that cannot be matched in the lexical context in which they appear. Unknown words are also treated as segmentation points, since the lexical transfer has no bilingual information for them and no structural transfer pattern is activated at all.

Once the SL (Occitan) corpus has been segmented, for each segment, all possible translations into TL (Catalan) according to every possible combination of disambiguations are obtained. Then, the likelihoods of these translations are computed through a Catalan trigram model trained from a 2-million-word raw-text Catalan corpus, and then normalized and used to estimate the HMM parameters as described in section 3.

We evaluated the evolution of the performance of the training method by updating the HMM parameters at every 1 000 words and testing the resulting part-of-speech tagger; this also helps in determining the amount of SL text required for the convergence.

Figure 2 shows the evolution of the word error rate (WER) when training the Occitan part-of-speech tagger from a 300 000-word raw-text Occitan corpus built from texts collected from the Internet. The results achieved when following the standard (unsupervised) Baum-Welch approach to train HMM-based part-of-speech taggers on the same corpus (no larger Occitan corpora was available to us in order to train with the Baum-Welch algorithm), and the results achieved when a TL model is used at translation time (instead of a SL part-of-speech tagger) to select always the most likely translation into TL (TLM-best) are given for comparison.

When reestimating the HMM parameters via the Baum-Welch algorithm, the log-likelihood of the training corpus was calculated after each iteration; the iterative reestimation process is finished when the difference between the log-likelihood of the last iteration and the previous one is below a certain threshold. Note that when training the HMM parameters via the Baum-Welch algorithm, the whole 300 000-word corpus is used, therefore the WER reported in figure 2 for the Baum-Welch algorithm is independent of the number of SL words in the horizontal axis.

The WER is calculated as the edit distance (Levenshtein, 1965) between the translation of an independent 10 079-word Occitan corpus performed by the MT system when embedding the part-of-speech tagger being evaluated, and its human-corrected MT into Catalan. WERs are calculated at the document level; additions, deletions and substitutions being equally weighted.

As can be seen in figure 2 our method does not need a large amount of SL text to converge and the translation performance is better than that achieved by the Baum-Welch algorithm. Moreover, the translation performance achieved by our method is even better than that achieved when translating using the TLM-best setup. Although the TLM-best setup might be though as giving the best result that can be achieved by our method, the results reported in figure 2 suggest that our method has some generalization capability that makes it able to produce better part-of-speech taggers for MT than it may be initially expected.

It must be mentioned that analogous re-

method is implemented inside package `apertium-tagger-training-tools` which is licensed under the GNU GPL license.

[7] The linguistic data for this language pair (package `apertium-o-ca-1.0.2`) can be freely downloaded from `http://apertium.sourceforge.net`

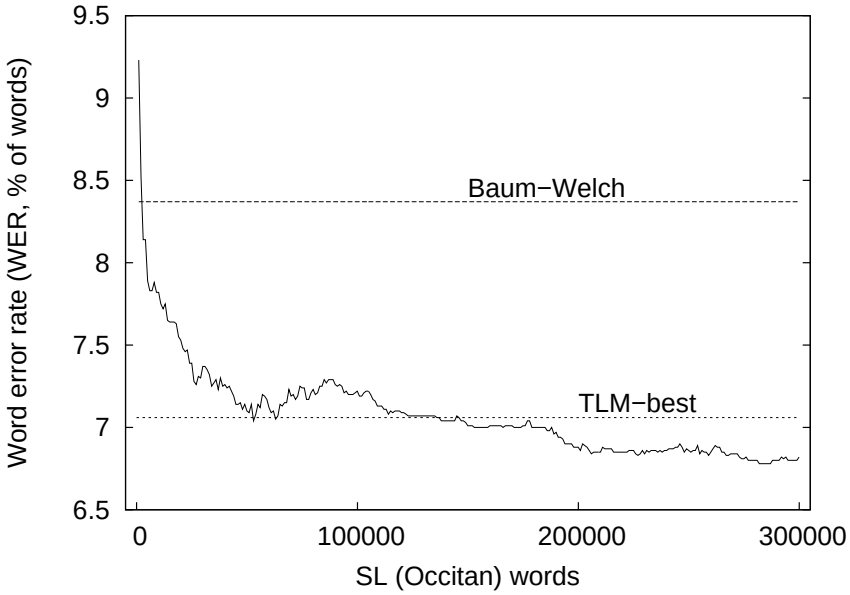

Figure 2: Evolution of the word error rate (WER) when training the (SL) Occitan part-of-speech tagger, Catalan being the target language (TL). WERs reported are calculated at the document level. Baum-Welch and TLM-best (see below) results are given for comparison; thus, they are independent of the number of SL words. TLM-best corresponds to the results achieved when a TL model is used at translation time (instead of a SL part-of-speech tagger) to select always the most likely translation into TL.

sults on the Spanish–Catalan language pair has revealed that, although the part-of-speech tagging accuracy is better when the HMM is trained in a supervised way from a tagged corpus, the translation performance of the MT system when embedding the supervisedly trained part-of-speech taggers is quite similar to that of using a part-of-speech tagger trained through the TL-driven training method.[8]

Concerning how the presented method behaves when the languages involved are less related than Occitan and Catalan, preliminary experiments on the French–Catalan language pair show results in agreement to those provided in this paper. Experiments on more unrelated languages pairs such as English–Catalan will be conducted in the near future.

5 Discussion

In this paper we have reviewed the use of target language (TL) information to train hidden-Markov-model (HMM)-based part-of-speech taggers to be used in machine translation (MT); furthermore, we have presented experiments done with the Occitan–Catalan language pair, a case study of a less-resourced language pair.

Our training method has been proven to be appropriate to train part-of-speech taggers for MT between less-resourced language pairs because, on the one hand, the amount of SL text needed is very small compared with common corpus sizes (millions of words) used by the Baum-Welch algorithm; and, on the other hand, because no new resources must be built (such as tagged corpora) to get translation performances comparable to those achieved when training from tagged corpora.

Finally, it must be pointed out that the resulting part-of-speech tagger is tuned to improve the translation quality and intended to be used as a module in a MT system; for this reason, it may give less accurate results as a general purpose part-of-speech tagger for other natural language processing applications.

Acknowledgements

Work funded by the Spanish Ministry of Education and Science through project TIN2006-15071-C03-01, by the Spanish Ministry of Education and Science and the European Social Fund through research grant BES-2004-4711, and by the Spanish Ministry of Indus-

[8] We plan to publish these results in the near future.

try, Tourism and Commerce through project FIT-350401-2006-5. The development of the Occitan–Catalan linguistic data was supported by the Generalitat de Catalunya.

References

Armentano-Oller, C., R.C. Carrasco, A.M. Corbí-Bellot, M.L. Forcada, M. Ginestí-Rosell, S. Ortiz-Rojas, J.A. Pérez-Ortiz, G. Ramírez-Sánchez, F. Sánchez-Martínez, and M.A. Scalco. 2006. Open-source Portuguese-Spanish machine translation. In *Computational Processing of the Portuguese Language, Proceedings of the 7th International Workshop on Computational Processing of Written and Spoken Portuguese, PROPOR 2006*, volume 3960 of *Lecture Notes in Computer Science*. Springer-Verlag, pages 50–59. (`http://www.dlsi.ua.es/~japerez/pub/pdf/propor2006.pdf`).

Armentano-Oller, C. and M.L. Forcada. 2006. Open-source machine translation between small languages: Catalan and Aranese Occitan. In *Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages)*, pages 51–54. (organized in conjunction with LREC 2006, `http://www.dlsi.ua.es/~mlf/docum/armentano06p2.pdf`).

Arnold, D., 2003. *Computers and Translation: A translator's guide*, chapter Why translation is difficult for computers, pages 119–142. Benjamins Translation Library. Edited by H. Somers.

Baum, L.E. 1972. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process. *Inequalities*, 3:1–8.

Carbonell, J., S. Klein, D. Miller, M. Steinbaum, T. Grassiany, and J. Frei. 2006. Context-based machine translation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, "Visions for the Future of Machine Translation"*, pages 19–28, August.

Corbí-Bellot, A.M., M.L. Forcada, S. Ortiz-Rojas, J.A. Pérez-Ortiz, G. Ramírez-Sánchez, F. Sánchez-Martínez, I. Alegria, A. Mayor, and K. Sarasola. 2005. An open-source shallow-transfer machine translation engine for the Romance languages of Spain. In *Proceedings of the 10th European Associtation for Machine Translation Conference*, pages 79–86, Budapest, Hungary. (`http://www.dlsi.ua.es/~mlf/docum/corbibellot05p.pdf`).

Cutting, D., J. Kupiec, J. Pedersen, and P. Sibun. 1992. A practical part-of-speech tagger. In *Third Conference on Applied Natural Language Processing. Association for Computational Linguistics. Proceedings of the Conference.*, pages 133–140, Trento, Italy.

Forcada, M.L. 2006. Open-source machine translation: an opportunity for minor languages. In *Proceedings of Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages)*. (`http://www.dlsi.ua.es/~mlf/docum/forcada06p2.pdf`).

Gale, W.A. and K.W. Church. 1990. Poor estimates of context are worse than none. In *Proceedings of a workshop on Speech and natural language*, pages 283–287. Morgan Kaufmann Publishers Inc.

Garrido-Alenda, A., M. L. Forcada, and R. C. Carrasco. 2002. Incremental construction and maintenance of morphological analysers based on augmented letter transducers. In *Proceedings of TMI 2002 (Theoretical and Methodological Issues in Machine Translation)*, pages 53–62.

Levenshtein, V.I. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848. English translation in Soviet Physics Doklady, 10(8):707-710, 1966.

Sánchez-Martínez, F., J.A. Pérez-Ortiz, and M.L. Forcada. 2004a. Cooperative unsupervised training of the part-of-speech taggers in a bidirectional machine translation system. In *Proceedings of TMI, The Tenth Conference on Theoretical and Methodological Issues in Machine Translation*, pages 135–144, October. (`http://www.dlsi.ua.es/~fsanchez/pub/pdf/sanchez04b.pdf`).

Sánchez-Martínez, F., J.A. Pérez-Ortiz, and M.L. Forcada. 2004b. Exploring the use of target-language information to train the part-of-speech tagger of machine translation systems. In *Advances in Natural Language Processing, Proceedings of*

4th International Conference EsTAL, volume 3230 of *Lecture Notes in Computer Science*. Springer-Verlag, pages 137–148. (http://www.dlsi.ua.es/~fsanchez/pub/pdf/sanchez04a.pdf).

Sánchez-Martínez, F., J.A. Pérez-Ortiz, and M.L. Forcada. 2006. Speeding up target-language driven part-of-speech tagger training for machine translation. In *Advances in Artificial Intelligence, Proceedings of the 5th Mexican International Conference on Artificial Intelligence*, volume 4293 of *Lecture Notes in Computer Science*. Springer-Verlag, pages 844–854. (http://www.dlsi.ua.es/~fsanchez/pub/pdf/sanchez06b.pdf).