Multiclass Classification for Morphology Generation in Statistical Machine Translation

Adolfo Hernández Huerta, Enric Monte Moreno, José B. Mariño Acebal

Department of Signal Theory and Communications
TALP Research Center (UPC)
Barcelona 08034, Spain
(adolfo.hernandez, enric.monte, jose.marino)@upc.edu

Abstract

We present a system for multiclass classification of simplified morphology of Spanish verbs within the framework of morphology generation for Statistical Machine Translation (SMT) from English into Spanish. In previous works it was proved that, when statistically translating from English into Spanish, the richness of morphology of the target language affects the translation models at training time by creating data sparseness. In order to determine the correct morphology of the Spanish translation we use a hierarchical set of classifiers through a Decision Directed Acyclic Graph (DDAG) structure, each decision-node operates with a classifier which is a Support Vector Machine (SVM). This structure is justified because it allows to introduce prior information about the difficulty of the task. The classification results are analyzed and commented.

Index Terms: morphology, machine learning, statistical machine translation

1. Introduction

Despite the fact that initially SMT systems ignored any linguistic analysis and worked at the surface level of word forms, there has been a growing effort to introduce linguistic knowledge into their statistical framework. It is clear that linguistic information has the potential to improve the performance of SMT systems, especially when limited amounts of parallel training data sets are available. However, incorporating linguistic, morphological, syntactic and semantic information into the statistical framework of SMT is a hard problem.

In particular, languages with rich morphology pose significant challenges for natural language processing. In highly inflected languages, the extensive use of inflection (to express agreement, gender, case, etc.), derivation, and composition leads to a huge vocabulary, forcing the translation model to learn different translation probability distributions for all inflected forms of nouns, adjectives or verbs, suffering thus from data sparseness. SMT systems estimated from parallel text are affected by this fact because it is impossible to train all forms from the training corpora. It is also important to point out that obtaining good performance in SMT systems when translating between languages with different morphological richness is a challenging task; especially when translating into a richer morphology language because the target language is represented by a larger vocabulary set, making decisions harder for SMT systems (e.g. higher perplexity in translation and target language models). However, different strategies have to be tackled when translating in the opposite direction (i.e. from a richer morphology language), where sparsity problems may arise in the source language (higher percentage of out-of-vocabulary (OOV) words, fewer translation examples for each input word, etc.).

Due to the above discussion, it is reasonable to find challenges related with morphology in SMT systems whose language pairs are English and any Romance-family language (Portuguese, Catalan, Galician, Italian, French, etc.) or pairs such as English and Arabic, Finnish, or German, to name a few. Then, the linguistic properties of the pair of languages and the translation direction pose severe limitations in most of the SMT tasks, such as word alignment and modeling. Several efforts are being done in the community to overcome such constraints by analyzing language specific problems and their impact on statistical translation as well as introducing some linguistic information in statistical models.

An important work was developed by Nießen and Ney [1], where, for a German-to-English task, several transformations of the source string are proposed, leading to an increased translation performance. These transformations include compound word separation, reordering of separated verb prefixes, and word mapping to word plus POS in order to distinguish articles from pronouns, among others. The same pair of languages and translation direction are used by Nießen and Ney [2] and Corston-Oliver and Gamon [3]. In the former, hierarchical lexicon models including base form and POS information are introduced, as well as other morphology-based data transformations. In the latter, inflectional normalization was achieved, leading to improvements in the perplexity of IBM translation models and reducing alignment errors. Aiming for a more general approach to deal with language-specific challenges, Koehn and Hoang [4] introduced factored translation models that can efficiently integrate morpho-syntactic information into phrase-based SMT. This framework adds in the model a vector of factors that represent different levels of annotation: word, lemma, POS, morphology, word class.

This work extends the research of de Gispert and Mariño [5], which dealt with the problem of the morphology derivation on Ngram-based Statistical Machine Translation (SMT) models from English into a morphology-rich language such as Spanish. It was shown that some Spanish morphology information could be introduced into simplified morphology translation hypotheses by means of an independent model. This approach is depicted in Figure 1. The translation system was proposed as a cascade-system integrated by a SMT system and a morphology generator. The SMT system consisted of simplified morphology translation models trained through a corpus with morphology simplification of words. From this study, several types of morphology simplification schemes were applied, depending

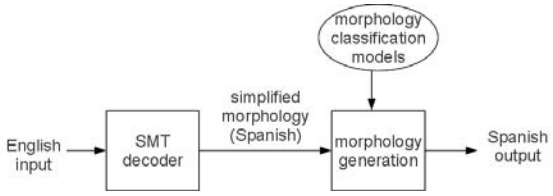

Fig. 1: *Translation architecture for SMT with morphology generation*

on which Part-Of-Speech category was modified (verbs, nouns, adjectives, etc.); it was concluded that the main source of potential improvement lied in verb form morphology. After the translation, Spanish morphology information is introduced into the simplified translation. For this purpose, a set of relevant features for each Spanish verb base form was defined in order to train statistical classifiers based on machine learning techniques, specifically through Adaboost, that used as base classifier a decision tree [6]. High accuracy scores were obtained when generating Spanish verb person, number and gender information, resulting in a significant improvement of final translation scores.

This paper is organized as follows. Section 2 briefly outlines the SMT system, the morphology generation and the multiclass classification approach. Section 3 reports and discusses the experimental results. Finally, Section 4 sums up the main conclusions from the paper.

2. System description

2.1. Ngram-based SMT system

The translation system implements a log-linear model in which a foreign language sentence $f_1^J = f_1, f_2 \ldots, f_J$ is translated into another language $e_1^I = e_1, e_2 \ldots, e_I$ by searching for the translation hypothesis $\hat{e}_1^I$ maximizing a log-linear combination of several feature models [7]:

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m \left(e_1^I, f_1^J \right) \right\} \qquad (1)$$

where the feature functions h_m refer to the system models and the set of λ_m refers to the weights corresponding to these models.

The core part of the system constructed in that way is a translation model, which is based on bilingual n-grams. It actually constitutes an Ngram-based language model of bilingual units (called tuples), which approximates the joint probability between the languages under consideration. The procedure of tuples extraction from a word-to-word alignment according to certain constraints is explained in detail by Mariño et al. [8].

The Ngram-based approach differs from the phrase based SMT mainly by distinct representating of the bilingual units defined by word alignment and using a higher order HMM of the translation process. While regular phrase-based SMT considers context only for phrase reordering but not for translation, the N-gram based approach conditions translation decisions on previous translation decisions.

The translation system, besides the bilingual translation model, which consists of a 4-gram LM of tuples with Kneser-Ney discounting (estimated with SRI Language Modeling Toolkit[1]), implements a log-linear combination of five additional feature models:

- a target language model (a 4-gram model of words, estimated with Kneser-Ney smoothing);
- a POS source language model (a 4-gram model of tags with Good-Turing discounting);
- a POS target language model (a 4-gram model of tags with Good-Turing discounting);
- a word bonus model, which is used to compensate the system's preference for short output sentences;
- a source-to-target lexicon model and a target-to-source lexicon model, these models use word-to-word IBM Model 1 probabilities [9] to estimate the lexical weights for each tuple in the translation table.

Decisions on the particular LM configuration and smoothing technique were taken on the minimal-perplexity and maximal-BLEU bases.

The decoder (called MARIE), an open source tool[2], implementing a beam search strategy was used in the translation system.

Given the development set and references, the log-linear combination of weights was adjusted using a simplex optimization method (with the optimization criteria of the highest BLEU score) and an n-best re-ranking just as described in http://www.statmt.org/jhuws/. This strategy allows for a faster and more efficient adjustment of model weights by means of a double-loop optimization, which provides significant reduction of the number of translations that should be carried out.

2.2. Morphology simplification

In the SMT system, after standard word alignment and tuple extraction, target language words (Spanish) are substituted with their simplified morphology forms. Then the bilingual N-gram translation model is estimated with these new tuples. The result is a standard bilingual model translating English into simplified morphology Spanish. The morphology simplification module produces a set of samples whose correct morphology is known, as they belong to the training corpus; these samples can be used to estimate morphology classification models.

In the case of simplification of information about person and number for Spanish verbs, the verb form, for instance, *'apoyen'* is transformed into *'VMSPpn[apoyar]'*, indicating simplified Part of Speech (POS) and base form. Under this simplification, the POS keeps information on word category (*'VM'*: Main Verb), mode and tense (*'SP'*: subjunctive, present), whereas *'p'* and *'n'* represent any person and number. Furthermore, as the correct person and number for this verb is known beforehand, it also serves as the class label during the training phase of the classifier that generates the morphology.

2.3. Multiclass classification

Morphology generation is implemented by means of classification models which, making use of a set of relevant features for each simplified morphology word and its context, generates its appropriate morphology. In order to tackle this task, we use a Decision Directed Acyclic Graph (DDAG), which combines many two-class classifiers into a multiclassification task. The use of this structure is justified because it allows to introduce

[1] http://www-speech.sri.com/projects/srilm/

[2] http://www.talp.cat/talp/index.php/ca/recursos/eines/marie

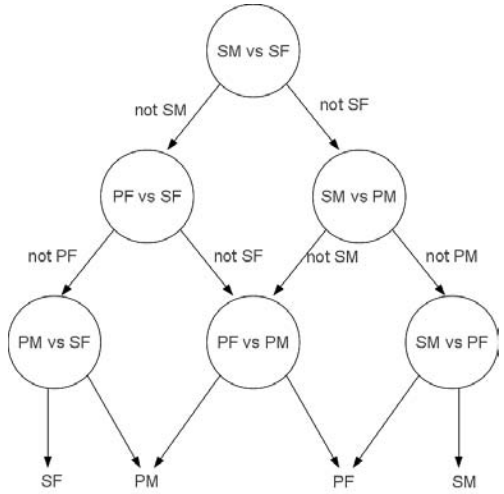

Fig. 2: *Decision DAG to find the best class out of four classes, it was the morphology related to gender and number*

EPPS corpus		sent.	words	vcb	avg.len.
train	Eng	1.40 M	39.29 M	121.07 k	28.02
	Spa		41.76 M	157.66 k	29.79
dev.	Eng	1996	58.63 k	6.54 k	29.37
test	Eng	1094	26.91 k	3.95 k	24.60

Tab. 1: *English-Spanish European Parliament corpus statistics*

information about the difficulty of the task. Instead of classifying one class versus all the others it does a pairwise comparison. As the set of classifiers are organized in a tree structure, the upper levels are assigned to the pairs of classes that have the lower error rate and also that the number of samples of each class is balanced (i.e. approximately the same number of examples per class).

The description of the structure is as follows. For an N-class problem, the DDAG contains N(N-1)/2 nodes, one for each pair of classes (one-vs-one classifier). A DAGSVM algorithm is proposed by Platt et al. [10], it proved to be superior to other multiclass SVM algorithms in both training and evaluation time. A DAGSVM places one-vs-one SVMs into the nodes of a DDAG. An example of a structure of the DDAG is shown in Figure 2.

3. Experiments

3.1. Database

This work was carried out on a large-data English-to-Spanish task, defined by a corpus containing official transcriptions of the European Parliament Plenary Sessions (EPPS), whose statistics are presented in Table 1. This corpus is available through the ACL-WMT evaluation campaign of year 2009[3].

3.2. Classification of Verb Forms

In order to generate morphology, two subcategories are distinguished: a) verb forms whose person and number (PN) information is missing (i.e. 1st person singular (1S), 2nd person singular (2S), and so on); b) verb forms whose number and gender (NG) is missing (i.e. past participle, which can also be regarded as adjective). Regarding the features, they were defined in lowercase text and extracted by a set of rules using word, POS tag, and base form information from both languages. Features take information from: a) bilingual model, i.e. current and previous tuples; b) target language, such as personal pronouns, presence of auxiliary verb 'haber' (for past participles), and verb form without 'tense' and 'mode' information; and c) source language, such as presence of English full verb form in the tuples, its POS tag, base form and related personal pronoun (if any, including reflexive pronouns, such as 'show them'), as well as presence of active or passive voice.

In the first case (PN), we trained the classifiers with 500k samples, while the second case (NG) the classifiers were trained with 300k samples. Finally, 10k samples were set apart for testing. The SVM$^{\text{Light}}$ algorithm [11] is used for training[4] .

In order to keep a balance in the one-vs-one classifiers as discussed in section 2.3, we decided to join clases for 2nd person and 3rd person singular (2S and 3S, respectively), as well

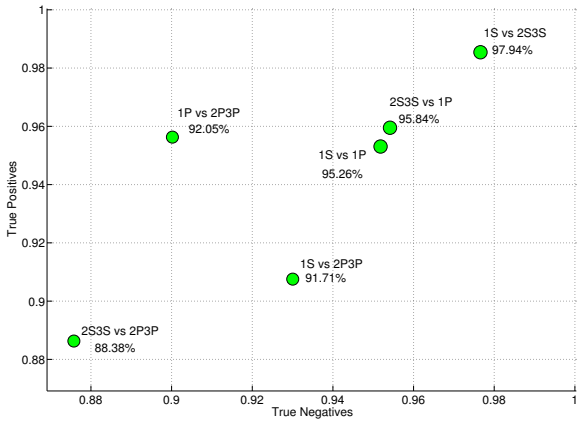

Fig. 3: *Spatial relationship between true positives, true negatives and accuracies for person and number*

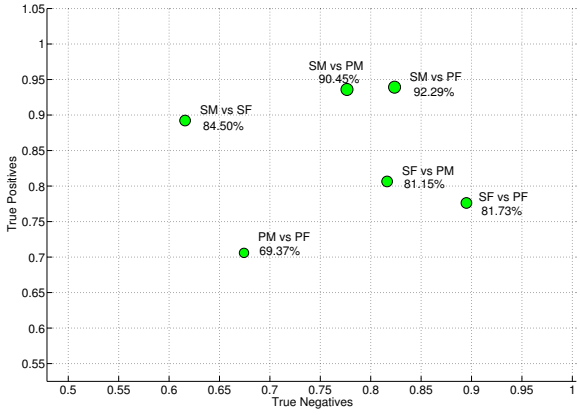

Fig. 4: *Spatial relationship between true positives, true negatives and accuracies for number and gender*

[3] http://www.statmt.org/wmt09

[4] http://svmlight.joachims.org/

classifier	accuracy
1S vs 2S3S	97.94%
1S vs 1P	95.26%
1S vs 2P3P	91.71%
2S3S vs 1P	95.84%
2S3S vs 2P3P	88.38%
1P vs 2P3P	92.05%

Tab. 2: *Classification accuracies on Spanish verb person and number morphology information*

classifier	accuracy
SM vs SF	84.50%
SM vs PM	90.45%
SM vs PF	92.29%
SF vs PM	81.15%
SF vs PF	81.73%
PM vs PF	69.37%

Tab. 3: *Classification accuracies on Spanish verb number and gender morphology information*

as 2n person and 3rd person plural (2P and 3P, respectively). This is due to the low number of training samples for 2S and 2P was rather low. Finally, in both subcategories, 4 classes were defined (i.e. 6 binary classifiers). Accuracies for each classifier alone are shown in Tables 2 and 3.

In general, accuracies to classify person and number are higher than accuracies for number and gender. Classifier related to PM-vs-PF shows the lowest accuracy. Figures 3 and 4 reflect our results; the nearer a classifier is to the upper right part of the figures, the better its performance. These figures allow us to compare the performance of the different one-vs-one classifiers.

Table 4 presents the accuracies obtained in the multiclassification task. Results for PN are more satisfactory than those for NG, it is clear that classifier PM-vs-PF requires bigger improvement.

4. Conclusions

In this paper we presented a system for multiclass classification of simplified morphology for SMT. The DDAG structure provides good accuracy results to classify person and number of Spanish verbs; however classification accuracies need to be improved to better generate gender and number. Future work to improve the performance of the classifiers will be directed to optimize the set of features used by each classifier and to improve the representation of the features through coding alternatives to reduce the dimensionality of the vectors.

5. Acknowledgements

This work has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 247762. This project was financed by the Ministry of Science and Innovation, ref. TEC2009-14094-C04-01. The first author is granted by a FI grant from the Catalan autonomous government.

morph. class.	DAG approach
Person & Number	85.70%
Number & Gender	72.31%

Tab. 4: *Classification accuracies of DDAG for PN and NG*

6. References

[1] S. Nießen and H. Ney, "Improving SMT quality with morpho-syntactic analysis," in *Proceedings of the 18th International Conference on Computational Linguistics*, Saarbrücken, Germany, July 2000, pp. 1081–1085.

[2] S. Niessen and H. Ney, "Statistical machine translation with scarce resources using morpho-syntactic information," *Computational Linguistics*, vol. 30, no. 2, pp. 182–204, Jun 2004.

[3] S. Corston-Oliver and M. Gamon, "Normalizing German and English inflectional morphology to improve statistical word alignment," in *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas*, October 2004, pp. 48–57.

[4] P. Koehn and H. Hoang, "Factored translation models," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 868–876. [Online]. Available: http://www.aclweb.org/anthology/D/D07/D07-1091

[5] A. de Gispert and J. B. Mariño, "On the impact of morphology in English to Spanish statistical MT," *Speech Communication*, vol. 50, no. 11-12, pp. 1034–1046, 2008.

[6] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," in *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, 1998, pp. 80–91. [Online]. Available: citeseer.nj.nec.com/schapire99improved.html

[7] P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," *Computational Linguistics*, vol. 16, no. 2, pp. 76–85, 1990.

[8] J. Mariño, R. Banchs, J. Crego, A. de Gispert, P. Lambert, J. Fonollosa, and M. Costa-jussà, "N-gram-based machine translation," *Computational Linguistics*, vol. 32, no. 4, pp. 527–549, 2006.

[9] F. J. Och and H. Ney, "The alignment template approach to statistical machine translation," *Computational Linguistics*, vol. 30, no. 4, pp. 417–449, 2004.

[10] J. C. Platt, N. Cristianini, and J. Shawe-taylor, "Large margin DAGs for multiclass classification," in *Advances in Neural Information Processing Systems*. MIT Press, 2000, pp. 547–553.

[11] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, 1999, ch. 11, pp. 169–184.