Combining Morphosyntactic Enriched Representation with n-best Reranking in Statistical Translation

H. Bonneau-Maynard, A. Allauzen, D. Déchelotte and H. Schwenk
Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{maynard,allauzen,dechelot,schwenk}@limsi.fr

Abstract

The purpose of this work is to explore the integration of morphosyntactic information *into* the translation model itself, by enriching words with their morphosyntactic categories. We investigate word disambiguation using morphosyntactic categories, n-best hypotheses reranking, and the combination of both methods with word or morphosyntactic n-gram language model reranking. Experiments are carried out on the English-to-Spanish translation task. Using the morphosyntactic language model alone does not results in any improvement in performance. However, combining morphosyntactic word disambiguation with a word based 4-gram language model results in a relative improvement in the BLEU score of 2.3% on the development set and 1.9% on the test set.

1 Introduction

Recent works in statistical machine translation (SMT) shows how phrase-based modeling (Och and Ney, 2000a; Koehn et al., 2003) significantly outperform the historical word-based modeling (Brown et al., 1993). Using phrases, i.e. sequences of words, as translation units allows the system to preserve local word order constraints and to improve the consistency of phrases during the translation process. Phrase-based models provide some sort of context information as opposed to word-based models. Training a phrase-based model typically requires aligning a parallel corpus, extracting phrases and scoring them using word and phrase counts. The derived statistics capture the structure of natural language to some extent, including implicit syntactic and semantic relations.

The output of a SMT system may be difficult to understand by humans, requiring re-ordering words to recover its syntactic structure. Modeling language generation as a word-based Markovian source (an n-gram language model) discards linguistic properties such as long term word dependency and word-order or phrase-order syntactic constraints. Therefore, explicit introduction of structure in the language models becomes a major and promising focus of attention.

However, as of today, it seems difficult to outperform a 4-gram word language model. Several studies have attempted to use morphosyntactic information (also known as part-of-speech or POS information) to improve translation. (Och et al., 2004) have explored many different feature functions. Reranking n-best lists using POS has also been explored by (Hasan et al., 2006). In (Kirchhoff and Yang, 2005), a factored language model using POS information showed similar performance to a 4-gram word language model. Syntax-based language models have also been investigated in (Charniak et al., 2003). All these studies use word phrases as translation units and POS information in just a post-processing step.

This paper explores the integration of morphosyntactic information *into* the translation model itself by enriching words with their morphosyntactic cat-

Proceedings of SSST, NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation, pages 65–71,
Rochester, New York, April 2007.

egories. The same idea has already been applied in (Hwang et al., 2007) to the Basic Travel Expression Corpus (BTEC). To our knowledge, this approach has not been evaluated on a large real-word translation problem. We report results on the TC-STAR task (public European Parliament Plenary Sessions translation). Furthermore, we propose to combine this approach with classical n-best list reranking. Experiments are carried out on the English-to-Spanish task using a system based on the publicly available *Moses* decoder.

This paper is organized as follows: In Section 2 we first describe the baseline statistical machine translation systems. Section 3 presents the considered task and the processing of the corpora. The experimental evaluation is summarized in section 4. The paper concludes with a discussion of future research directions.

2 System Description

The goal of statistical machine translation is to produce a target sentence $\mathbf{e}$ from a source sentence $\mathbf{f}$. Among all possible target language sentences the one with the highest probability is chosen. The use of a maximum entropy approach simplifies the introduction of several additional models explaining the translation process:

$$\begin{aligned} \mathbf{e}^* &= \arg\max \Pr(\mathbf{e}|\mathbf{f}) \\ &= \arg\max_e \{\exp(\sum_i \lambda_i h_i(\mathbf{e}, \mathbf{f}))\} \quad (1) \end{aligned}$$

where the feature functions h_i are the system models characterizing the translation process, and the coefficients λ_i act as weights.

2.1 Moses decoder

Moses[1] is an open-source, state-of-the-art phrase-based decoder. It implements an efficient beam-search algorithm. Scripts are also provided to train a phrase-based model. The popular Giza++ (Och and Ney, 2000b) tool is used to align the parallel corpora.

The baseline system uses 8 feature functions h_i, namely phrase translation probabilities in both directions, lexical translation probabilities in both directions, a distortion feature, a word and a phrase penalty and a trigram target language model. Additional features can be added, as described in the following sections. The weights λ_i are typically optimized so as to maximize a scoring function on a development set (Och and Ney, 2002).

The moses decoder can output n-best lists, producing either distinct target sentences or not (as different segmentations may lead to the same sentence). In this work, distinct sentences were always used.

These n-best lists can be rescored using higher order language models (word- or syntactic-based). There are two ways to carry out the rescoring: one, by replacing the language model score or by adding a new feature function; two, by performing a log-linear interpolation of the language model used for decoding and the new language model. This latter approach was used in all the experiments described in this paper. The set of weights is systematically re-optimized using the algorithm presented below.

2.2 Weight optimization

A common criterion to optimize the coefficients of the log-linear combination of feature functions is to maximize the BLEU score (Papineni et al., 2002) on a development set (Och and Ney, 2002). For this purpose, the public numerical optimization tool *Condor* (Berghen and Bersini, 2005) is integrated in the following iterative algorithm:

0. Using good general purpose weights, the Moses decoder is used to generate 1000-best lists.
1. The 1000-best lists are reranked using the current set of weights.
2. The current hypothesis is extracted and scored.
3. This BLEU score is passed to *Condor*, which either computes a new set of weights (the algorithm then proceeds to step 1) or detects that a local maxima has been reached and the algorithm stops iterating.

The solution is usually found after about 100 iterations. It is stressed that the n-best lists are generated only once and that the whole tuning operates only on the n-best lists.

[1] http://www.statmt.org/moses/

English: I$_{PP}$ declare$_{VVP}$ resumed$_{VVD}$ the$_{DT}$ session$_{NN}$ of$_{IN}$ the$_{DT}$ European$_{NP}$ Parliament$_{NP}$

Spanish: declaro$_{VLfin}$ reanudado$_{VLadj}$ el$_{ART}$ perodo$_{NC}$ de$_{PREP}$ sesiones$_{NC}$ del$_{PDEL}$ Parlamento$_{NC}$ Europeo$_{ADJ}$

Figure 1: Example of POS-tag enriched bi-text used to train the translation models

2.3 POS disambiguation

It is well-known that syntactic structures vary greatly across languages. Spanish, for example, can be considered as a highly inflectional language, whereas inflection plays only a marginal role in English.

POS language models can be used to rerank the translation hypothesis, but this requires tagging the n-best lists generated by the SMT system. This can be difficult since POS taggers are not well suited for ill-formed or incorrect sentences. Finding a method in which morphosyntactic information is used directly in the translation model could help overcome this drawback but also takes account for the syntactic specificities of both source and target languages. It seems likely that the morphosyntactic information of each word will be useful to encode linguistic characteristics, resulting in a sort of word disambiguation by considering its morphosyntactic category. Therefore, in this work we investigate a translation model which enriches every word with its syntactic category. The enriched translation units are a combination of the original word and the POS tag, as shown in Figure 1. The translation system takes a sequence of enriched units as inputs and outputs. This implies that the test data must be POS tagged before translation. Likewise, the POS tags in the enriched output are removed at the end of the process to provide the final translation hypothesis which contain only a word sequence. This approach also allows to carry out a n-best reranking step using either a word-based or a POS-based language model.

3 Task, corpus and tools

The experimental results reported in this article were obtained in the framework of an international evaluation organized by the European TC-STAR project[2] in February 2006. This project is envisaged as a long-term effort to advance research in all core technologies for speech-to-speech translation.

The main goal of this evaluation is to translate public European Parliament Plenary Sessions (EPPS). The training material consists of the summary edited by the European Parliament in several languages, which is also known as the Final Text Editions (Gollan et al., 2005). These texts were aligned at the sentence level and they are used to train the statistical translation models (see Table 1 for some statistics).

	Spanish	*English*
	Whole parallel corpus	
Sentence Pairs	1.2M	
Total # Words	34.1M	32.7M
Vocabulary size	129k	74k
	Sentence length ≤ 40	
Sentence Pairs	0.91M	
Total # Words	18.5M	18.0M
Word vocabulary	104k	71k
POS vocabulary	69	59
Enriched units vocab.	115k	77.6k

Table 1: Statistics of the parallel texts used to train the statistical machine translation system.

Three different conditions are considered in the TC-STAR evaluation: translation of the Final Text Edition (*text*), translation of the transcriptions of the acoustic development data (*verbatim*) and translation of speech recognizer output (*ASR*). Here we only consider the *verbatim* condition, translating from English to Spanish. For this task, the development and test data consists of about 30k words. The test data is partially collected in the Spanish parliament. This results in a small mismatch between development and test data. Two reference translations are provided. The scoring is case sensitive and includes punctuation symbols.

[2] http://www.tc-star.org/

3.1 Text normalization

The training data used for normalization differs significantly from the development and test data. The Final Text Edition corpus follows common orthographic rules (for instance, the first letter of the word following a full stop or a column is capitalized) and represents most of the dates, quantities, article references and other numbers in digits. Thus the text had to be "true-cased" and all numbers were verbalized using in-house language-specific tools. Numbers are not tagged as such at this stage; this is entirely left to the POS tagger.

3.2 Translation model training corpus

Long sentences (more than 40 words) greatly slow down the training process, especially at the alignment step with Giza++. As shown in Figure 2, the histogram of the length of Spanish sentences in the training corpus decreases steadily after a length of 20 to 25 words, and English sentences exhibit a similar behavior. Suppressing long sentences from the corpus reduces the number of aligned sentences by roughly 25% (see Table 1) but speeds the whole training procedure by a factor of 3. The impact on performance is discussed in the next section.

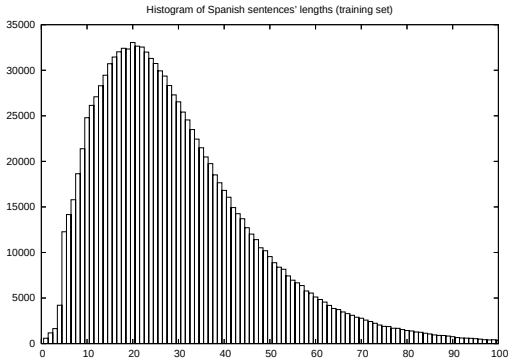

Figure 2: Histogram of the sentence length (Spanish part of the parallel corpus).

3.3 Language model training corpus

In the experiments reported below, a trigram word language model is used during decoding. This model is trained on the Spanish part of the parallel corpus using only sentences shorter than 40 words (total of 18.5M of language model training data). Second pass language models were trained on all available monolingual data (34.1M words).

3.4 Tools

POS tagging was performed with the *TreeTagger* (Schmid, 1994). This software provides resources for both of the considered languages and it is freely available. *TreeTagger* is a Markovian tagger that uses decision trees to estimate trigram transition probabilities. The English version is trained on the *PENN treebank* corpus[3] and the Spanish version on the *CRATER* corpus.[4]

Language models are built using the SRI-LM toolkit (Stolcke, 2002). Modified Knesser-Ney discounting was used for all models. In (Goodman, 2001), a systematic description and comparison of the usual smoothing methods is reported. *Modified Knesser-Ney* discounting appears to be the most efficient method.

4 Experiments and Results

Two baseline English-to-Spanish translation models were created with Moses. The first model was trained on the whole parallel text – note that sentences with more than 100 words are excluded by Giza++. The second model was trained on the corpus using only sentences with at most 40 words. The BLEU score on the development set using good general purpose weights is 48.0 for the first model and 47.0 for the second. Because training on the whole bi-text is much slower, we decided to perform our experiments on the bi-texts restricted to the "short" sentences.

4.1 Language model generation

The reranking experiments presented below use the following language models trained on the Spanish part of the whole training corpus:

- word language models,
- POS language model,
- POS language model, with a stop list used to remove the 100 most frequent words (POS-stop100 LM),
- language model of enriched units.

[3] http://www.cis.upenn.edu/ treebank

[4] http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html

English :	you will be aware President that over the last few sessions in Strasbourg. ..
Baseline:	usted sabe que el Presidente *durante los últimos sesiones* en Estrasburgo ...
Enriched units:	usted sabe que el Presidente *en los últimos períodos de sesiones* en Estrasburgo ...
English :	... in this house there might be some recognition ...
Baseline:	... en esta asamblea *no puede ser un cierto reconocimiento* ...
Enriched units:	... en esta asamblea *existe un cierto reconocimiento* ...

Figure 3: Comparative translations using the baseline word system and the enriched unit system.

For each of these four models, various orders were tested ($n = 3, 4, 5$), but in this paper we only report those orders that yielded the greatest improvements. POS language models were obtained by first extracting POS sequences from the previously POS-tagged training corpus and then by estimating standard back-off language models.

As shown in Table 1, the vocabulary size of the word language model is 104k for Spanish and 74k for English. The number of POS is small: 69 for Spanish and 59 for English. We emphasize that the tagset provided by *TreeTagger* does include neither gender nor number distinction. The vocabulary size of the enriched-unit language model is 115k for Spanish and 77.6k for English. The syntactical ambiguity of words is low: the mean ambiguity ratio is 1.14 for Spanish and 1.12 for English.

4.2 Reranking the word n-best lists

The results concerning reranking experiments of the n-best lists provided by the translation model based on *words as units* are summarized in Table 2. The baseline result, with trigram word LM reranking, gives a BLEU score of 47.0 (1rst row). From the n-best lists provided by this translation model, we compared reranking performances with different target language models. As observed in the literature, an improvement can be obtained by reranking with a 4-gram word language model (47.0 $\rightarrow$ 47.5, 2d row). By post-tagging this n-best list, a POS language model reranking can be performed. However, reranking with a 5-gram POS language model alone does not give any improvement from the baseline (BLEU score of 46.9, 3rd row). This result corresponds to known work in the literature (Kirchhoff and Yang, 2005; Hasan et al., 2006), when using POS only as a post-processing step during reranking. As suggested in section 2.3, this lack of performance can be due to the fact that the tagger is not able to provide a usefull tagging of sentences included in the n-best lists. This observation is also available when reranking of the word n-best is done with a language model based on enriched units (BLEU score of 47.6, not reported in Table 2).

4.3 POS disambiguation and reranking

The results concerning reranking experiments of the n-best lists provided by the translation model based on *enriched units* are summarized in Table 3. Using a trigram language model of enriched translation units leads to a BLEU score of 47.4, a 0.4 increase over the baseline presented in section 4.2. Figure 3 shows comparative translation examples from the baseline and the enriched translation systems. In the first example, the baseline system outputs "*durante los últimos sesiones*" where the enriched translation system produces "*en los últimos períodos de sesiones*", a better translation that may be attributed to the introduction of the masculine word "*períodos*", allowing the system to build a syntactically correct sentence. In the second example, the syntactical error "*no puede ser un cierto reconocimiento*" produced by the baseline system induces an incorrect meaning of the sentence, whereas the enriched translation system hypothesis "*existe un cierto reconocimiento*" is both syntactically and semantically correct.

Reranking the enriched n-best with POS language models (either with or without a stop list) does not seem to be efficient (0.3 BLEU increasing with the POS-stop100 language model).

A better improvement is obtained when reranking is performed with the 4-gram word language model. This results in a BLEU score of 47.9, corresponding to a 0.9 improvement over the word baseline. It is interesting to observe that reranking a n-best list

	Dev.	Test
3g word LM baseline	47.0	46.0
4g word LM reranking	**47.5**	**46.5**
5g POS reranking	46.9	46.1

Table 2: BLEU scores using words as translation units.

obtained with a translation model based on enriched units with a word LM results in better performances than a enriched units LM reranking of a n-best list obtained with a translation model based on words.

The last two rows of Table 3 give results when combining word and POS language models to rerank the enriched n-best lists. In both cases, 10 features are used for reranking (8 Moses features + word language model probability + POS language model probability). The best result is obtained by combining the 5-gram word language model with the 5-gram POS-stop100 language model. In that case, the best BLEU score is observed (48.1), with a 2.3% relative increase over the trigram word baseline.

4.4 Results on the test set

The results on the test set are given in the second column of Tables 2 and 3. Although the enriched translation system is only 0.1 BLEU over the baseline system (46.0 $\rightarrow$ 46.1) when using a trigram language model, the best condition observed on the development set (word and POS-stop100 LMs reranking) results in a 46.8 BLEU score, corresponding to a 0.8 increasing.

It can be observed that rescoring with a 4-gram word language model leads to same score resulting in a 1.9% relative increase over the trigram word baseline.

5 Conclusion and future work

Combining word language model reranking of n-best lists based on syntactically enriched units seems to produce more consistent hypotheses. Using enriched translation units results in a relative 2.3% improvement in BLEU on the development set and 1.9% on the test over the trigram baseline. Over a standard translation model with 4-gram rescoring, the enriched unit translation model leads to an absolute increase in BLEU score of 0.4 both on the development and the test sets. These first results are encouraging enough to further investigate the integration of syntactic information in the translation model itself, rather than to restrict it to the post-processing pass. As follow-up experiments, it is planned to include gender and number information in the tagset, as well as the word stems to the enriched units.

	Dev.	Test
3g enriched units LM baseline	47.4	46.1
4g enriched units LM reranking	47.8	46.8
4g word LM reranking	47.9	46.9
5g POS LM reranking	47.5	46.2
5g POS-stop100 LM reranking	47.7	46.3
word + POS LMs reranking	47.9	46.9
word + POS-stop100 LMs rerank.	**48.1**	**46.8**

Table 3: BLEU scores using enriched translation units.

This work should be considered as preliminary experiments for the investigation of factored translation models, which Moses is able to handle. POS factorization is indeed a way to add some generalization capability to the enriched translation models.

6 Acknowledgments

This work has been partially funded by the European Union under the integrated project TC-STAR (IST-2002-FP6-506738), and by the French Government under the project INSTAR (ANR JCJC06_143038).

We would like to thanks Marc Ferras for his help concerning the Spanish language.

References

Frank Vanden Berghen and Hugues Bersini. 2005. CONDOR, a new parallel, constrained extension of powell's UOBYQA algorithm: Experimental results and comparison with the DFO algorithm. *Journal of Computational and Applied Mathematics*, 181:157–175.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

E. Charniak, K. Knight, and K. Yamada. 2003. Syntax-based language models for machine translation. In *Proceedings of MT Summit IX*.

C. Gollan, M. Bisani, S. Kanthak, R. Schlueter, and H̃. Ney. 2005. Cross domain automatic transcription on

the TC-STAR epps corpus. In *Proceedings of ICASSP 2005*.

Joshua T. Goodman. 2001. A bit of progress in language modeling. *Computer Speech and Language*, 15(4):403–434, October.

S. Hasan, O. Bender, and H. Ney. 2006. Reranking translation hypothesis using structural properties. In *Proceedings of EACL 2006*.

Y.S. Hwang, A. Finch, and Y. Sasaki. 2007. Improving statistical machine translation using shallow linguistic knoledge. *to be published in Computer, Speech and Language*.

Katrin Kirchhoff and Mei Yang. 2005. Improved language modeling for statistical machine translation. In *Proceedings of ACL '05 workshop on Building and Using Parallel Text*, pages 125–128.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology Conference 2003 (HLT-NAACL 2003)*, Edmonton, Canada, May.

Franz Josef Och and Hermann Ney. 2000a. Improved statistical alignment models. In *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447, Hongkong, China, October.

Franz Josef Och and Hermann Ney. 2000b. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447, Hong Kong, China, October.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL 2002*, pages 295–302.

F.-J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, and D. Radev. 2004. A smorgasbord of features for statistical machine translation. In *NAACL*, pages 161–168.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, University of Pennsylvania.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, September.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of ICSLP*, pages II: 901–904.