Exploiting Word Transformation in Statistical Machine Translation from Spanish to English

Deepa Gupta and Marcello **Federico**

ITC-irst-Centro per la Ricerca Scientifica e Tecnologica, Via Sommarive 18, 38050 Povo (Trento) Italy

{gupta|federico}@itc.it

Abstract

This paper investigates the use of morpho-syntactic information to reduce data-sparseness in statistical machine translation from Spanish to English. In particular, word-alignment training is performed by applying different word transformations using lemmas and stems. It has been observed that stem-based training is better than lemma-based training when up to 1 million running words of data are used. In this paper a new word-alignment training technique is proposed by exploiting syntactically motivated constraints to the parallel data. Preliminary experimental results show that stem-based training with syntactically motivated constraints gives significant improvement in translation performance. Finally, a technique to reduce the impact of out-of-vocabulary words is discussed. The considered task is the translation of Plenary Sessions of the European Parliament.

1 Introduction

In this work we investigate the use of morpho-syntactic information to improve performance of a phrase-based statistical machine translation (SMT) system. The considered task is the translation of Plenary Sessions of the European Parliament (EPPS) from Spanish to English.

Recent results on this topic are reported in Popović & Ney (2004), Popović & Ney (2005), and Popović, Vilar, Ney, Jovicic, & Saric (2005) where stem-suffix and lemma-Part-of-Speech (POS) information were used for translating German, Serbian, Spanish into English and vice versa. Moreover, in Gispert (2005) lemma-POS information restricted to verbs was used in a translation task from Spanish to English. As a difference with respect to previous work, we investigate word transformations which are blindly applied to all word categories and under different data sparseness conditions. Moreover, morpho-syntactic knowledge is only applied during training of the SMT system, and is not directly exploited during translation. As an exception, stems of Spanish words have been also used for translating out-of-vocabulary (OOV) words in the test set.

The organization of the paper is as follows. Section 2 introduces different lexicon transformations. Section 3 gives details about the data and the SMT system. Sections 4 and 5 describe techniques to reduce data-sparseness in SMT and discuss experimental results. Section 6 explains a technique to reduce the impact of OOV words. Finally, Section 7 contains concluding remarks.

2 Lexicon Transformations

We applied two kinds of lexicon transformations, lemmatization and stemming, both on Spanish and English.

Lemmatization: we employed the FreeLing[1] tool, a POS tagger for English and Spanish, which also provides the base form or lemma for each input word. Figure 1 shows an example of the output of FreeLing.

[1] http://garraf.epsevg.upc.es/freeling

Input sentence: ¿ hay alguna observación ?		
Original word	Lemma/base form	POS tag+ morpho-attributes
¿	¿	Fia
hay	haber	VAIP3S0
alguna	alguno	DI0FS0
observación	observación	NCFS000
?	?	Fit

Figure 1: Example of FreeLing output on Spanish. The complex tag "VAIP3S0" for word *hay* means: verb(V), auxiliary verb(A), indicative mode(I), present(P), third person (3), singular number (S) and no gender(0)

Spanish Sentence	
Original :	¿ hay alguna observación ?
Stemmed:	¿ hay algun observ ?
English Sentence	
Original:	are there any comments ?
Stemmed:	ar there ani comment ?

Figure 2: Example of stemmed texts.

Stemming: we used the Snowball stemmer[2] for English and Spanish. It is based on the Porter's algorithm which truncates and replaces suffixes of words. Figure 2 shows the output of the stemmer for Spanish and English.

3 EPPS Task

Data

Data for the task was prepared by and is available from the TC-STAR project[3]. Table 1 provides statistics about increasing portions of the parallel corpus used for training, and about the test data. The average length of sentences is 16.7 for Spanish and 15.8 for English. From Table 1 it results that Spanish shows higher rates of singleton words which is an indicator of data sparseness. The percentage of out-of-vocabulary (OOV) words on the test data is also reported.

SMT system

Given a string f in the source language, the ITC-irst SMT system (Cettolo, Federico, Bertoldi, Cattoni, & Chen, 2005), looks for the target string e maximizing the posterior probability $\Pr(\mathbf{e},\mathrm{a} \mid \mathrm{f})$ over all possible word alignments a. The conditional distribution is computed with the log-linear model:

$$p_{\lambda}(\mathbf{e},\mathrm{a} \mid \mathrm{f}) \quad \propto \quad \exp\{\sum_{r=1}^{R} \lambda_r h_r(\mathbf{e},\mathrm{f},\mathrm{a})\},$$

where $h_r(\mathbf{e},\mathrm{f},\mathrm{a}), r = 1 \ldots R$ are real valued feature functions.

The log-linear model is used to score translation hypotheses $(\mathbf{e},\mathrm{a})$ built in terms of strings of phrases, which are simple n-grams of words. The phrase-based translation (Koehn, Och, & Marcu, 2003) process works as follows (Federico & Bertoldi, 2005). At each step, a target phrase is added to the translation whose corresponding source phrase within f is identified through three random quantities: the *fertility* which establishes its length; the *permutation* which sets its first position; the *tablet* which tells its word string. Notice that target phrases might have fertility equal to zero, hence they do not translate any source word. Moreover, untranslated words in f are also modelled through some random variables.

The above process is modelled with eight feature functions (Cettolo et al., 2005) whose parameters are either estimated from data (e.g. target language models, phrase-based lexicon models) or empirically fixed (e.g. permutation models). While feature functions exploit statistics extracted from monolingual or word-aligned texts from the training data, the scaling factors λ of the log-linear model are estimated on the development data by applying a *minimum error training* procedure (Och, 2004).

[2] http://www.snowball.tartarus.org/

[3] http://tc-star.itc.it.

Training data									
			words			lemmas		stems	
	sent.	run. words	vocab.	sings.	OOV	vocab.	sings.	vocab.	sings.
Spanish	7.5k	126,761	7,470	1,723	8.39%	4,603	891	3,980	650
English		118,348	5,402	1,019	-	4,204	724	3,676	587
Spanish	15k	253,513	10,220	1,593	6.14%	6,134	824	5,108	571
English		235,241	7,263	981	-	5,597	728	4,810	579
Spanish	30k	502,556	16,862	4,248	3.82%	10,002	2,352	7,760	1,437
English		475,107	11,549	2482	-	8,914	1,864	7,543	1,514
Spanish	60k	1,006,675	25,486	7,966	2.50%	15,256	4,780	11,021	2,744
English		956,006	16,613	4,311	-	12,863	3,338	10,698	2,623
Spanish	120k	2,007,490	36,162	11,344	1.83%	22,313	7,394	15,168	3,949
English		1,908,891	22,717	5,910	-	17,809	4,727	14,705	3,739
Test set									
Spanish	sent.	run. words	vocab.	sings.	avg. sent. len.		max. sent. len.		
	1,008	27,683	3.649	2,030	27.5		112		

Table 1: Statistics of different training corpus sizes (number of sentence pairs) for the EPPS task, out-of-vocabulary (OOV) rate in the test data and statistics of the test data.

The phrase-based lexicon model is computed from a parallel corpus provided with word-alignments in both directions, i.e. from source to target positions, and vice-versa. Word alignments are computed with the GIZA++ toolkit (Och & Ney, 2003). Translation pairs of phrases are extracted in a way to preserve the original word alignments (Cettolo et al., 2005).

The target language model exploits a 3-gram language models estimated on 39.4 million words from the EPPS corpus. Finally, the search algorithm that computes the most probable translation is implemented with a beam-search algorithm explained in Federico & Bertoldi (2005).

The following section addresses data sparseness issues and investigates the use of word transformations.

4 Reduction of Data Sparseness

Spanish is a morphologically richer language than English. However, all inflected forms of Spanish are not relevant for translation into English. For instance, the adjective "*bonito*" (beautiful/pretty) has four inflected forms ("*bonita*", "*bonitas*", "*bonito*", "*bonitos*") according to the gender and number of the noun it modifies. This is not the case of English adjectives, which only have one form. Therefore, it might be possible that all inflected forms of Spanish adjectives are not required for translation. Similar cases are possible to a limited extent with other words also, such as nouns, verbs etc.

In this work we investigate if better word alignment is achieved by transforming words in the training data either with lemmas or stems. When 7.5K sentences pair of training data are used, Spanish vocabulary reduces approximately by 38% using lemmas from 7,470 to 4,603 and by 47% using stems from 7,470 to 3,980. Table 1 shows reduction in the vocabulary size for other training data sizes.

Once the training corpus is transformed by using either lemmas or stems, word-alignment is performed in both directions. After, lemmas and stems are replaced again with words and phrase-extraction is performed. To evaluate translation quality we used well known translation measures: BLUE, NIST, Word Error Rate (WER) and Position-independent Error Rate (PER). Automatic scores were computed by exploiting two reference translations for each test sentence.

Table 2 reports experimental results on

Alignment	Train	BLEU(%)	NIST	WER(%)	PER(%)
Word-based	7.5k	30.36	7.24	50.55	41.15
Lemma-based		31.76	7.52	49.39	39.69
Stem-based		32.33	7.62	49.17	39.15
Word-based	15k	35.39	8.06	47.71	37.42
Lemma-based		36.12	8.22	47.25	36.74
Stem-based		36.26	8.27	47.01	36.41
Word-based	30k	39.65	8.62	45.16	34.54
Lemma-based		39.77	8.75	45.10	34.33
Stem-based		40.03	8.77	44.84	34.21
Word-based	60k	42.37	9.08	43.50	32.81
Lemma-based		42.21	9.05	43.56	33.22
Stem-based		42.37	9.07	43.41	33.01
Word-based	120k	44.37	9.29	42.08	31.65
Lemma-based		43.88	9.24	42.08	31.65
Stem-based		43.98	9.23	42.59	32.10

Table 2: Translation with different word-alignments and amounts of sentence pairs.

all different training corpus sizes. For instance, performing lemma-based alignment on 7.5K parallel sentences gives relative improvements of 4.6% BLUE, 3.8% NIST, 2.29% WER and 3.55% PER. Using stems instead gives improvements of 6.4% BLUE, 5.2% NIST, 2.73% WER and 4.86% PER. In general, stem-based alignment gives more consistent improvements than lemma-based training. However, after about 60K sentence pairs, both stem- and lemma-based alignments do not improve translation scores with respect to word-based alignment. The reason could be that beyond a given amount of training data, the reduction of data sparseness does not compensate for the loss of information caused by the word transformation.

To compensate this loss of information, we investigated word alignment training augumented with syntactically motivated constraints.The better results obtained with stem-based alignment have motivated us to use this method in the subsequent experiments.

5 Alignment with Syntactic Constraints

One difficulty of word-alignment is related to the length of sentences, which determines the space of possible word alignments. It is also known from common practice that alignment quality can be improved by adding a bilingual dictionary to the training data. In fact, from a Bayesian point of view, a dictionary can be considered as prior knowledge supplied to the alignment model, which somehow constraints the possible word-to-word mappings.

However, finding reliable constraints is not always possible. In this work, we have investigated the use of syntactic knowledge to generate constraints from the training corpus itself.

Noun-part Translation Constraints

One peculiarity of Spanish-English translation is the almost preservation of noun parts and prepositions. In other words, a Spanish word which is either a determiner, noun, adjective, possessive pronoun, or article will be very likely translated into an English word whose POS belongs to the same group. Evidence of this property was observed in the EPPS data, too. This property has been used to select phrase-pairs from the parallel corpus to be used as prior knowledge for the alignment model.

For each sentence pair in the training data, the sub-strings containing only words of the noun-preposition group were extracted, from both source and target sides. Hence, alignment was performed and most

Alignment +constraints	Train	BLEU(%)	NIST	WER(%)	PER(%)
Word-based	7.5k	30.74	7.30	50.11	40.59
Stem-based		32.26	7.60	48.80	38.93
Word-based	15k	35.66	8.11	47.04	36.94
Stem-based		36.81	8.34	45.99	35.68
Word-based	30k	40.23	8.77	44.29	34.06
Stem-based		40.78	8.87	43.78	33.45
Word-based	60k	43.37	9.15	42.47	32.06
Stem-based		43.15	9.17	42.44	32.30
Word-based	120k	44.41	9.29	41.7	31.34
Stem-based		44.92	9.33	41.44	31.36

Table 3: Experimental results of word-alignment training with constraints.

reliable (frequent) phrase pairs were extracted. Notice that resulting phrase-pairs are not necessarily meaningful, given that words in-between can be missing.

Such frequent phrase-pairs were then added to the training data as an additional parallel corpus. In particular, each entry was weighted by its frequency and all entries were scaled by an empirically set factor.

Alignment was again performed on the augmented data according to the word-based and stem-based modalities. Notice that final phrase-pairs used to estimate the translation model were only built on the alignments of the original training corpus.

Table 3 shows translation results with the two alignment methods. It appears that the use of constraints improves translation scores in almost all data-sparseness conditions. Only in the case of 7.5K sentences pairs, there is a marginal reduction in the BLUE and NIST scores when stem-based alignment is performed. Most significant improvements occur in the two largest training sets with the stem-based alignment. A possible explanation is that the introduction of constraints reduces the alignment ambiguity, and compensates for the loss of information caused by word stemming.

So far we have discussed methods to improve word alignment training. The subsequent section discusses a method to reduce the impact of OOV words during translation.

6 Translation of OOV Words by Stems

During translation, we replace OOV words in the test set with their stems. The rationale is that stems of OOV words could correspond to stems of observed words, for which a correct or almost correct translation is indeed available. For instance, this should work for Spanish adjectives, for which all inflections correspond to the same English form. However, in this investigation we applied this concept to all words indistinctly.

Phrase-pair statistics used to build the translation model were augmented with Spanish-stem to English-word translation pairs extracted from the aligned training corpus. In particular, stem-based alignment was employed, with all English stems replaced with the original words. In this way, during decoding, the search algorithm is able to resort on stem-to-word translations for the out-of-vocabulary words found in the test sentences. We can see in Figure 3 that for 7.5K sentence pairs training data, the lexicon augmentation gives a relative OOV rate reduction of 22.2% (from 8.39% to 6.53%), while for 120K sentence pairs, the reduction is 49.2% (from 1.83% to 0.93%).

Figure 4 compares performance, in terms of BLEU score, of the main SMT systems proposed in this work, namely: the baseline trained on word-based alignment (baseline), the system using stem-based alignment (stem), the one also using constraints,

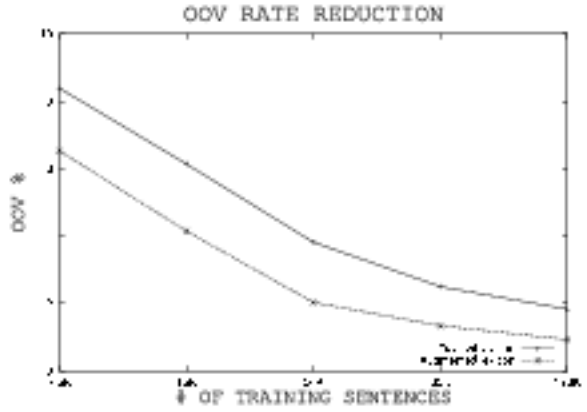

Figure 3: OOV rate after lexicon augmentation.

Figure 4: BLEU score after lexicon augmentation.

and the system also using lexicon augmentation. Performance improvements were obtained by lexicon augmentation under all training data conditions, but the 120k sentence corpus which indeed has the lowest OOV rate.

For the 15K sentence-pair condition, the final BLEU improvement is of 5.91% (from 35.39% to 37.48%). From Figure 4), it can be clearly observed that the improvements of all proposed methods are almost additive, and more effective as data sparseness increases.

7 Conclusion

In this paper, we have systematically investigated the impact of word-based, lemma-based and stem-based word alignment training on translation performance under different training data sizes. We have shown that stem-based alignment training gives better results than lemma-based and word-based training. We have also improved word alignment training by exploiting syntactically motivated constraints. Our results showed consistent improvement in translation performance when stem-based alignment training is applied. Finally, we have proposed a method to cope with the translation of OOV words found in the source string. OOV words occurring in the input string are replaced with their stems, and the phrase-based lexicon model is augmented with stem-to-word translation pairs extracted from the aligned corpus.

Ongoing research is exploring syntactic constraints with other syntactic groups and their combinations in word alignment training.

8 Acknowledgements

This work has been funded by the European Union under the integrated project TC-STAR- Technology and Corpora for Speech to Speech Translation-(IST-2002-FP6-506738. http://www.tc-star.org)

References

Cettolo, M., Federico, M., Bertoldi, N., Cattoni, R., & Chen, B. (2005). A look inside the itc-irst smt system. In *Proceedings of the 10th Machine Translation Summit* (pp. 451–457). Phuket, Thailand.

Federico, M., & Bertoldi, N. (2005). A word-to-phrase statistical translation model. *ACM Transactions on Speech and Language Processing*, *2*(2), 1–24.

Gispert, A. de. (2005). Phrase linguistic classification and generalization for improving statistical machine translation. In *ACL Student Research Workshop* (pp. 67–72). Ann Arbor, Michigan.

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 3rd Conference of the human language technology conference/ North American Chapter of the ACL* (pp. 127–133). Edmonton, Canada.

Och, F. J. (2004). Minimum error rate training in statistical machine translation. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*. Sapporo, Japan.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19 – 51.

Popović, M., & Ney, H. (2004). Towards the use of word stems and suffixes for statistical machine translation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation* (pp. 1585 – 1588). Lisbon, Portugal.

Popović, M., & Ney, H. (2005). Exploiting phrasal lexica and additional morpho-syntactic language resources for statistical machine translation with scarce training data. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation* (pp. 212 – 218). Budapest, Hungary.

Popović, M., Vilar, D., Ney, H., Jovicic, S., & Saric, Z. (2005). Augmenting a small parallel text with morpho-syntactic language resources for serbian-english statistical machine translation. In *Proceedings of the 43rd Meeting of the Association for Computational Linguistics* (pp. 41 – 48). Ann Arbor, Michigan.