Vocabulary Extension via PoS Information for SMT

Germán Sanchis, Joan Andreu Sánchez
Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
Camino de Vera, s/n. 46022 Valencia, Spain
{gsanchis,jandreu}@dsic.upv.es

Abstract

One of the weaknesses of the so-called phrase based translation models is that they carry out a blind extraction of the phrase translation table, i.e., they do not take into account the linguistic information which is inherent to every language. On the other hand, Part of Speech (PoS) tagging is a problem that, nowadays, presents a pretty mature state of the art, obtaining error rates of almost 2%. Because of this, the use of automatically PoS-tagged corpora in Statistical Machine Translation (SMT) with the purpose of incorporating syntactical knowledge and enhancing the results obtained by state of the art SMT systems seems quite natural. In this work, we present results obtained on the EuroParl corpus by creating an extended vocabulary composed of the regular words and their PoS tags concatenated to them.

1 Introduction

Machine Translation (MT) is a research field of great importance in the European Community, where language plurality implies both a very important cultural richness and not negligible obstacle towards building a unified Europe. Because of this, a growing interest on MT has been shown both by politicians and research groups, which become more and more specialised in this field. Although the language plurality problem can be seen as a more global problem, reaching in fact world wide, in this paper we will be focusing on European languages, due to the vast amount of free data which is available for them.

Moreover, Statistical Machine Translation (SMT) systems are receiving an increasing importance in the last years. In the tasks they have been trained on, SMT systems are able to deliver similar translation quality than rule-based machine translation systems, with the benefit of requiring little human effort when adapting to new language pairs, whenever suitable corpora are available.

(Brown et al., 1993) established what is considered nowadays as the mathematical background of modern SMT, defining the machine translation problem as follows: given a sentence s from a certain source language, an adequate sentence $\hat{t}$ that maximises the posterior probability is to be found. This leads to the following formula:

$$\hat{t} = \underset{t}{\operatorname{argmax}}\, p(t|s)$$

Applying the Bayes theorem on this definition, one can easily reach the next formula

$$\hat{t} = \underset{t}{\operatorname{argmax}}\, \frac{p(t) \cdot p(s|t)}{p(s)}$$

and, since we are maximising over t, the denominator can be neglected, arriving to

$$\hat{t} = \underset{t}{\operatorname{argmax}}\, p(t) \cdot p(s|t)$$

where $p(t|s)$ has been decomposed into two different probabilities: the *statistical language model* of the target language $p(t)$ and *the (inverse) translation model* $p(s|t)$.

Although it might seem odd to model the probability of the source sentence given the target sentence, this decomposition has a very intuitive interpretation: the translation model $p(s|t)$ will account for the possible word relations which can be established between input and output language, whereas the language model $p(t)$ will ensure that the output sentence is a well-formed sentence belonging to the target language.

Recently, there have been several efforts, coming from various research groups, to incorporate syntactic information into SMT systems (Kirchhoff et al., 2006; Popović and Ney, 2006a). More specifically, Part of Speech (PoS) tags have been used with the purpose of reordering the input or output sentence and obtaining a monotonous translation (Popović and Ney, 2006b).

In this context, we will be exploring the usefulness of including PoS information within the surface form (i.e. words) in each language, with the purpose of performing a sort of disambiguation over words which cannot be differentiated otherwise. We will be applying this extension on Moses (Koehn et al., 2007b), a phrase based (PB) SMT system.

Similar work was performed throughout the JHU Summer Workshop 2006 (Koehn et al., 2007a), when Moses was first built. In this work, however, factored translation models were used, and results on only a fraction of EuroParl were reported.

The rest of this work is structured as follows: first, in section 2, we will make a brief overview of the state of the art in PoS tagging. In section 3, we will review briefly phrase based SMT systems. In the next section, the experimental setup we carried out is detailed, and the translation results obtained are presented in section 5. Lastly, section 6 presents the conclusions we arrived to.

2 Part-Of-Speech Tagging

As is usual in many fields of Pattern Recognition and Language Modelling, the first PoS taggers were rule based systems (Greene and Rubin, 1971; Brill, 1992). However, the tasks where these systems could be applied belonged to a very restricted field, although their use was enough general to enable them to build tagged corpora, which were later on revised by human experts. These corpora were the key towards developing new, more efficient taggers.

More recently, the statistical framework gained a lot of importance, mainly because of the easiness with which the statistical models could be applied to new tasks. In fact, state-of-the-art PoS tagging is still driven by Hidden Markov Models (HMM), which were first applied by (Church, 1988), and later on by (Brants, 2000), who developed a tagger which still now belongs to the state of the art.

Within this framework, the hidden states represent the tags, whereas the observables are the words in the original corpus. Hence, transition probabilities depend of the origin and target states, i.e., tag pairs. On the other hand, observables only depend on the PoS tag assigned in the emission state. Formally, as defined by (Brants, 2000):

$$\underset{l_1 \ldots l_T}{\operatorname{argmax}} \left[\prod_{i=1}^{T} p(l_i|l_{i-1}, l_{i-2}) \cdot p(w_i|l_i) \right] \cdot p(l_{T+1}|l_T)$$

where $w_1 \ldots w_T$ is a sequence of words for length T and $l_1 \ldots l_T$ are elements of the set of PoS tags. The tags l_{-1}, l_0, l_{T+1} are tags indicating the beginning and the end of the sequence and are added to the set of tags for coherence purposes, but also because their inclusion implies a slight performance increase.

Moreover, Brants introduced a smoothing technique based on unigram, bigram and trigram interpolation, obtaining the following formula for the probability of a trigram:

$$\begin{aligned} p(l_3|l_1, l_2) \quad = \quad & \lambda_1 \hat{p}(l_3) + \lambda_2 \hat{p}(l_3|l_2) + \\ & \lambda_3 \hat{p}(l_3|l_1, l_2) \end{aligned} \qquad (1)$$

where $\hat{p}$ are the maximum likelihood estimations of the probabilities, and λ_n represents the weights of each one of the n-grams, obeying the restriction $\lambda_1 + \lambda_2 + \lambda_3 = 1$, so that p will remain a probability distribution.

In this work, we will be using the TnT Tagger (Brants, 2000) for tagging the German–English corpus and the FreeLing (Asterias et al., 2006) for tagging the Spanish–English corpus. Both these taggers are HMM based. Although PoS tagging is a monolingual problem and the English side of both parallel corpora could be tagged with the same toolkit, we did not do so because we took advantage of data we already had available, and both taggers present similar precision rates of over 97% (Brants, 2000; Asterias et al., 2006).

3 Phrase-based models

In the last years, phrase based (PB) models (Tomas and Casacuberta, 2001; Marcu and Wong, 2002; Zens et al., 2002; Zens and Ney, 2004) have proved to provide a very efficient framework for MT. Computing the translation probability of a given *phrase*, i.e. a sequence of words, and hence introducing information about context, these SMT systems seem to have mostly outperformed single-word models, quickly evolving into the predominant technology in the state of the art (Koehn and Monz, 2006a).

3.1 The model

The derivation of PB models stems from the concept of bilingual segmentation, i.e. sequences of source words and sequences of target words. It is assumed that only segments of contiguous words are considered, the number of source segments being equal to the number of target segments (say K) and each source segment being aligned with only one target segment and vice versa.

Let I and J be the lengths of t and s respectively[1]. Then, the bilingual segmentation is formalised through two segmentation functions: μ for the target segmentation ($\mu_1^K : \mu_k \in \{1, 2, \ldots, I\}$, $0 < \mu_1 \leq \mu_2 \leq \ldots \leq \mu_k = I$) and γ for the source segmentation ($\gamma_1^K : \gamma_k \in \{1, 2, \ldots, J\}$, $0 < \gamma_1 \leq \gamma_2 \leq \ldots \leq \gamma_k = J$). The alignment between segments is introduced through the alignment function α ($\alpha_1^K : \alpha_k \in \{1, 2, \ldots, K\}, \alpha(k) = \alpha(k')$ iff $k = k'$).

By assuming that all possible segmentations of s in K phrases and all possible segmentations of t in K phrases have the same probability independent of K, then $p(s|t)$ can be written as:

$$p(s|t) \propto \sum_K \sum_{\mu_1^K} \sum_{\gamma_1^K} \sum_{\alpha_1^K} \prod_{k=1}^{K} p(\alpha_k|\ \alpha_{k-1}) \cdot p(s_{\gamma_{\alpha_k-1}+1}^{\gamma_{\alpha_k}} | t_{\mu_{k-1}+1}^{\mu_k}) \quad (2)$$

where the distortion model $p(\alpha_k|\ \alpha_{k-1})$ (the probability that the target segment k is aligned with the source segment α_k) is usually assumed to depend only on the previous alignment α_{k-1} (first order model).

3.2 Learning phrase-based models

Ultimately, when learning a PB model, the purpose is to compute a *phrase translation table*, in the form

$$\{(s_j \ldots s_{j'}), (t_i \ldots t_{i'}), p(s_j \ldots s_{j'} | t_i \ldots t_{i'})\}$$

where the first term represents the input (source) phrase, the second term represents the output (target) phrase and the last term is the probability assigned by the model to the given phrase pair.

In the last years, a wide variety of techniques to produce PB models have been researched and implemented (Koehn et al., 2003). Firstly, a direct learning of the parameters of the equation $p(s_j^{j'} | t_i^{i'})$ was proposed (Tomas and Casacuberta, 2001; Marcu and Wong, 2002). At the same time, heuristics for extracting all possible segmentations coherent with a word-aligned corpus (Zens et al., 2002), where the alignments were learnt by means of the GIZA++ toolkit (Och and

[1] Following a notation used in (Brown et al., 1993), a sequence of the form $z_i, \ldots, z_j$ is denoted as z_i^j. For some positive integers N and M, the image of a function $f : \{1, 2, \ldots, N\} \rightarrow \{1, 2, \ldots, M\}$ for n is denoted as f_n, and all the possible values of the function as f_1^N

Ney, 2003), were also proposed. Other approaches have been suggested, exploring more linguistically motivated techniques (Sánchez and Benedí, 2006; Watanabe et al., 2003). In this paper, we report experiments using the heuristic, (word) alignment-based phrase extraction algorithm.

3.3 Decoding in phrase-based models

Once a SMT system has been trained, a decoding algorithm is needed. Different search strategies have been suggested to define the way in which the search space is organised. Some authors (Ortiz et al., 2003; Germann et al., 2001) have proposed the use of an $A^\star$ algorithm, which adopts a *best-first* strategy that uses a stack (priority-queue) in order to organise the search space. On the other hand, a *depth-first* strategy was also suggested in (Berger et al., 1996), using a set of stacks to perform the search.

4 Experimental setup

In this section we will be describing the Europarl corpus (Koehn, 2005) on which we performed our experiments, how it is structured, how we added PoS information to the system built, and how this information affects the language model and vocabulary sizes.

4.1 The Europarl corpus

The Europarl corpus (Koehn, 2005) is built from the proceedings of the European Parliament, which are published on the web, and was acquired in 11 different languages. However, in this work we will only focus on the German–English and Spanish–English corpus, due to the fact that it is much easier to find good PoS taggers for these languages.

For our experiments, we used the second version of this corpus, which is the one described in (Koehn, 2005) and the one that was used in the 2006 Workshop on Machine Translation of the NAACL (Koehn and Monz, 2006b). This corpus is divided into four separate sets: one for training, one for development, one for test and another test set which was the one used in the workshop for the final evaluation. This test set will be referred to as "Test", whereas the test set provided for evaluation purposes outside the final evaluation will be referred to as "Devtest". It must be noted that the Test set included a surprise out-of-domain subset, and hence the translation quality on this set will be significantly lower.

Since the original corpus is not sentence-aligned, and not every English sentence has its corresponding translation in German and Spanish (or vice-versa), two different corpora are obtained while constructing the German–English and Spanish–English parallel bilingual corpora. The characteristics of these corpora can be seen in Table 1.

It seems important to point that the average length of the sentences in German is always shorter than the average mean of the sentences in English, and the sentences in English are as well longer than the ones in Spanish. Moreover, the vocabulary size in German is more than 2,5 times bigger than the English vocabulary. This is due to the agglutinative nature of German, that has the ability of building compound words from simple words. For example, "Nachttisch" comes from the words "Nacht" and "Tisch" and means, literally "nighttable". This grants German an enormous lexical richness, but hinders the training of MT systems that involve German, either as source or target language. In addition, the fact that the average sentence length in the training subsets is much shorter than in the other sets is because in the cited workshop the training set was restricted to sentences with a maximum length of 40 words, whereas the other three subsets did not have this restriction.

Since the translations in the corpus have been written by a big number of different human translators, a same sentence may be translated in several different ways, all of them correct. This fact increases the difficulty of the corpus, and can be seen in the number of different pairs that constitute the training set, which is very similar to the total number of pairs. An example is the English sentence "*We shall now proceed to vote.*". It appears translated both as "*Se procede a la*

Table 1: Characteristics of the German–English and Spanish–English Europarl corpus

		German	English	Spanish	English
Training	Sentences	751088		730740	
	Different pairs	735792		715615	
	Running words	15257871	16052702	15725136	15222505
	Vocabulary size	195291	65889	102886	64123
	Average length	20.3	21.4	21.5	20.8
Development	Sentences	2000		2000	
	Running words	55147	58655	60628	58655
	Average length	27.6	29.3	30.3	29.3
	Out of vocabulary	432	125	208	127
Devtest	Sentences	2000	2000	2000	2000
	Running words	54260	57951	60332	57951
	Average length	27.1	29.0	30.2	29.0
	Out of vocabulary	377	127	207	125
Test	Sentences	3064	3064	3064	3064
	Running words	82477	85232	91730	85232
	Average length	26.9	27.8	29.9	27.8
	Out of vocabulary	1020	488	470	502

Table 2: Perplexity of the various corpus subsets with 3-grams and 5-grams.

		3-gram	5-gram
Dev	German	127.6	148.6
	English	74.6	89.9
	Spanish	74.2	89.0
Devtest	German	128.8	149.8
	English	73.7	88.9
	Spanish	75.3	90.6
Test	German	199.7	221.1
	English	118.5	134.5
	Spanish	103.2	117.9

votación.", which is quite a faithful translation, and "*El debate queda cerrado.*", which means "the debate is now closed". Although these two Spanish sentences are clearly different, one can clearly imagine a scenario where both translations would fit.

In the shared task of the NAACL06 Workshop on Statistical Machine Translation, the baseline system used 3-grams as language model, whereas in the shared task of the ACL07 Workshop, which used a newer and somewhat bigger version of the Europarl corpus, the baseline system was constructed with a language model consisting on 5-grams. Since we will be performing experiments both with 5-grams and with 3-grams, the perplexity of the various subsets of the corpus are shown in Table 2. These language models were computed with the SRILM (Stolcke, 2002) toolkit, applying interpolation with the Kneser-Ney discount.

4.2 Preparing the system

Before training the translation models, we PoS tagged all the subsets of the two corpora, obtaining a tagged bilingual corpus. Then, we concatenated the PoS tag to each one of the words, obtaining an extended vocabulary and producing two new different *"languages"*. Although the PoS taggers used have very high success rates, the fact of learning a translation model that involving PoS tags introduces noise in the system, and the error rates of the PoS tagger must affect the final translation quality. Nevertheless, we expect that the benefit obtained will be higher than the error introduced.

Given that for translating we will also need a target language model, we trained three new language models, one for each of the new *"languages"* that was produced by adding the PoS

Table 3: Perplexity of the various corpus subsets with concatenated PoS tags.

		3-gram	5-gram
Dev	German^PoS	129.9	151.1
	English^PoS	77.0	89.9
	Spanish^PoS	74.0	89.0
Devtest	German^PoS	130.9	152.0
	English^PoS	76.1	88.9
	Spanish^PoS	75.1	90.4
Test	German^PoS	202.7	223.7
	English^PoS	124.5	134.5
	Spanish^PoS	102.9	117.7

tags. Their with respect to the different subsets of the corpus is shown in table 3. It can be seen that the perplexity does not suffer an important variation by introducing the PoS tags. This seems encouraging, since it implies that adding the PoS information does not necessarily mean that the language model will be worse. However, it must also be taken into account that the vocabulary sizes do increase significantly: in the case of German, the size increases from 195291 to 212929, in the case of Spanish from 102886 to 109634 and in the case of English from 65889 to 81436, in the German–English subcorpus, and from 64123 to 79229 in the Spanish–English subcorpus. This means an increment of about 10% for German, 5% for Spanish and 22% for English. The fact that it is in English where the vocabulary size is most increased can be explained because of the relatively small vocabulary size that the original English corpus has: since there are fewer words, each word is bound to have, in average, a higher number of different syntactic functions, and hence will be assigned to a wider range of different PoS tags.

5 Translation Experiments

For our translation experiments we used the Moses toolkit (Koehn et al., 2007b). This toolkit involves the estimation of four different translation models, which are in turn combined in a log-linear fashion by adjusting a weight for each of them by means of the MERT (Och, 2003) procedure. For this purpose, a held-out corpus was used, namely the "Development" subset described in section 4.1.

Following previous works in SMT, and for comparability purposes, we will be evaluating our system with BLEU (Papineni et al., 2001) and WER. BLEU measures the precision of unigrams, bigrams, trigrams and 4-grams with respect to a set of reference translations, with a penalty for too short sentences. The WER criterion computes the minimum number of editions (substitutions, insertions and deletions) needed to convert the translated sentence into the sentence considered as ground truth. WER is a pessimistic measure when applied to MT.

Once the different corpus subsets had been tagged, we trained three different translation models.

The first one, which we used as baseline, was trained by applying the Moses toolkit directly. The second one was trained with the extended vocabulary corpus, using the extended words throughout the whole training and translation (decoding) process. Finally, a third translation model was learnt by using the extended vocabulary only to obtain the word alignments, necessary for the phrase-extraction algorithm to obtain phrases. The results can be seen in table 4. In all cases, we used a 5-gram language model, which is the one used as baseline for the 2007 Workshop in Machine Translation of the ACL.

In this table, the column "word^PoS" shows the results for the second experimental setup described above. The last column presents the results obtained by only using the extended vocabulary for alignment purposes.

Unfortunately, in the case of "word^PoS" almost all the results obtained are slightly (although not significantly) worse than those obtained with the baseline system. In the case of "pos-align", most of the results obtained improved by some tenths the baseline, except for the case of English→Spanish. On the other hand, adding PoS information seems to perform slightly better on the *test* set, where out-of-domain sentences were added. However,

Table 4: Translation scores when extending the vocabulary with the PoS tags.

		baseline		word^PoS		pos-align	
pair	subset	WER	BLEU	WER	BLEU	WER	BLEU
Es-En	Devtest	57.7	31.6	57.8	31.5	57.5	31.7
	Test	57.8	30.6	58.1	30.3	57.5	30.8
En-Es	Devtest	58.4	31.3	58.7	31.1	58.6	31.0
	Test	57.5	30.3	57.7	30.2	57.6	30.1
De-En	Devtest	65.5	26.2	65.5	26.2	65.0	26.3
	Test	68.1	23.7	68.7	23.7	67.5	24.1
En-De	Devtest	71.6	18.8	71.3	18.9	71.3	18.9
	Test	72.5	16.4	72.6	16.4	72.5	16.5

these slight improvements are not statistically significant.

Only as a small experiment, we checked what would the situation be if the language model used was a 3-gram instead of a 5-gram. In this case, and for the pair German→English, the score was boosted by 1.4 BLEU points on the *devtest* subset, from a 24.55 baseline score to a 25.95 obtained in the "word^PoS" setting. Quite interestingly, the score obtained in this setting is almost the same (just two tenths less) than the one obtained with a 5-gram. Hence, PoS information might be more useful in a task where the amount of data available is lower.

6 Conclusions

The results shown in this paper are discouraging in the sense that they seem to imply that adding PoS-tag information does not yield significant improvements on the quality of the final translation produced.

However, this might be so in the case of the EuroParl corpus, where a fairly big amount of data is available. Nevertheless, the use of PoS-tag information could be explored in tasks where the amount of training data is sparser. As future work, we plan to investigate this.

Acknowledgements

This work has been partially supported by the EC (FEDER) and the Spanish MEC under grant TIN2006-15694-CO2-01, by the MEC scholarship AP2005-4023 and by the Universidad Politécnica de Valencia with the ILELA project.

References

J. Asterias, B. Casas, E. Comelles, M. González, L. Padró, and M. Padró. 2006. Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy.

A.L. Berger, P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, J.R. Gillet, A.S. Kehler, and R.L. Mercer. 1996. Language translation apparatus and method of using context-based translation models. In *United States Patent 5510981.*

T. Brants. 2000. TnT - a statistical part-of-speech tagger. In *Proceedings of the Sixth ANLP*, Seattle, WA.

Eric Brill. 1992. A simple rule-based part of speech tagger. In *Proceedings of the Third ANLP*, Trento, Italy.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of machine translation. In *Computational Linguistics*, volume 19, pages 263–311, June.

K.W. Church. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the 1st ANLP*, pages 136–143, ACL.

U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada. 2001. Fast decoding and optimal decoding for machine translation. In *Proceeding of the 39th. Annual Meeting of the ACL*, pages 228–235, Toulouse, France.

B.B. Greene and G.M. Rubin. 1971. Automated grammatical tagging of English. In *Technical*

Report, Department of Linguistics, Brown University.

K. Kirchhoff, M. Yang, and K. Duh. 2006. Statistical machine translation of parliamentary proceedings using morpho-syntactic knowledge. In *TC-STAR Workshop on Speech-to-Speech Translation*, pages 57–62, Barcelona, Spain, June.

P. Koehn and C. Monz. 2006a. Manual and automatic evaluation of machine translation between european languages. In *Proceedings of the NAACL 2006, Workshop on SMT*, pages 102–121, New York City.

P. Koehn and C. Monz, editors. 2006b. *Proceedings on the Workshop on Statistical Machine Translation*. Association for Computational Linguistics, New York City, June.

P. Koehn, F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conf. of the NAACL on Human Language Technology*, volume 1, pages 48–54, Edmonton, Canada.

P. Koehn, M. Federico, W. Shen, N. Bertoldi, O. Bojar, C. Callison-Burch, B. Cowan, C. Dyer, H. Hoang, R. Zens, A. Constantin, C. Corbett Moran, and E. Herbst. 2007a. Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding. In *Final Report of the 2006 Language Engineering Workshop*, John Hopkins University, Center for Speech and Language Processing.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007b. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007*, Prague, Czech Republic, June.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.

D. Marcu and W. Wong. 2002. Joint probability model for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP02)*, Pennsylvania, Philadelphia, USA.

F.J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. In *Computational Linguistics*, volume 29, pages 19–51.

F.J. Och. 2003. Minimum error rate training for statistical machine translation. In *ACL 2003: Proc. of the 41st Annual Meeting of the ACL*, Sapporo, Japan, July.

D. Ortiz, I. García-Varea, and F. Casacuberta. 2003. An empirical comparison of stack-based decoding algorithms for statistical machine translation. In *New Advance in Computer Vision, Springer-Verlag, Lecture Notes in Computer Science, 1st Iberian Conference on Pattern Recongnition and Image Analysis (IbPRIA2003)*, Mallorca, Spain.

Papineni, A. Kishore, S. Roukos, T. Ward, and W. Jing Zhu. 2001. Bleu: A method for automatic evaluation of machine translation. In *Technical Report RC22176 (W0109-022)*, IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY.

M. Popović and H. Ney. 2006a. Error analysis of verb inflections in spanish translation output. In *TC-STAR Workshop on Speech-to-Speech Translation*, pages 99–103, Barcelona, Spain, June.

M. Popović and H. Ney. 2006b. Pos-based word reorderings for statistical machine translation. In *5th International Conference on Language Resources and Evaluation (LREC)*, pages 1278–1283, Genoa, Italy, May.

J.A. Sánchez and J.M. Benedí. 2006. Stochastic inversion transduction grammars for obtaining word phrases for phrase-based statistical machine translation. In *Proceedings of the Workshop on SMT*, pages 130–133, New York City.

A. Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 901–904.

J. Tomas and F. Casacuberta. 2001. Monotone statistical translation using word groups. In *Proceedings of the Machine Translation Summit VIII*, pages 357–361, Santiago de Compostela, Spain.

T. Watanabe, E. Sumita, and H.G. Okuno. 2003. Chunk-based statistical translation. In *Proceedings of the 41st. Annual Meeting of the ACL*, Sapporo, Japan.

R. Zens and H. Ney. 2004. Improvements in phrase-based statistical machine translation. In *Proceedings of the Human Language Technology Conference (HLT-NAACL)*, pages 257–264, Boston, USA.

R. Zens, F.J. Och, and H. Ney. 2002. Phrase-based statistical machine translation. In *Advances in artificial intelligence. 25. Annual German Conference on AI. Lecture Notes in Computer Science*, volume 2479, pages 18–32.