The TALP Ngram-based SMT System for IWSLT 2006

Josep M. Crego, Adrià de Gispert, Patrik Lambert,
Maxim Khalilov, Marta R. Costa-jussà, José B. Mariño,
Rafael Banchs and José A.R. Fonollosa

TALP Research Center
Universitat Politècnica de Catalunya, Barcelona
{jmcrego|agispert|lambert|khalilov|mruiz|canton|rbanchs|adrian}@gps.tsc.upc.edu

Abstract

This paper describes **TALPtuples**, the 2006 Ngram-based statistical machine translation system developed at the TALP Research Center of the UPC (Universitat Politècnica de Catalunya) in Barcelona. Emphasis is put on improvements and extensions of the system of previous years, being highlighted and empirically compared. Mainly, these include a novel and much more efficient word ordering strategy based on reordering patterns, a linguistically-guided tuple segmentation criterion and improved optimization procedures.

The paper provides details of this system participation in the third International Workshop on Spoken Language Translation (IWSLT) held in Kyoto, Japan in November 2006. Results on four translation directions are reported, namely from Arabic, Chinese, Italian and Japanese into English for the open data track, thoroughly explaining all language-related preprocessing and optimization schemes.

1. Introduction

Rooted in the Finite-State Transducers approach to SMT [1, 2] and estimating a joint-probability model between the source and the target languages, Ngram-based SMT has proved to be a very competitive alternative to phrase-based and other state-of-the-art systems in previous evaluation campaigns, as shown in [3]. This is specially true when dealing with pairs of languages with a relatively similar word order [4, 5].

Given the language pairs involved in this year's evaluation, efforts have been focused on improving the word reordering strategies for Ngram-based SMT. Specifically, a novel reordering strategy based on extending the search graph with automatically-extracted reordering patterns is explored. Results are very promising while keeping computational expenses at a similar level of monotone search. Additionally, a novel tuple segmentation strategy based on the entropy of Part-Of-Speech distributions was used with slight improvements in model estimation.

This paper is organized as follows. Section 2 briefly reviews last year's system, including tuple definition and extraction, translation model and feature functions, decoding tool and optimization criterion. Section 3 delves into the word ordering problem, by contrasting the constrained reordered search from previous years with the novel strategy based on reordering patterns. Section 4 focuses on tuple segmentation strategies, and contrasts the criterion on IBM model 1 probabilities from 2005 with a novel criterion based on Part-Of-Speech entropy distributions.

Later on, Section 5 reports on all experiments carried out from Arabic, Chinese, Italian and Japanese into English for IWSLT 2006. Finally, Section 6 sums up the main conclusions from the paper and discusses future research lines.

2. 2005 system review

The TALP Ngram-based SMT system performs a log-linear combination of a translation model and additional feature functions (see further details in [6, 7]). In contrast to phrase-based models, our translation model is estimated as a standard n-gram model of a bilingual language expressed in *tuples*. This way it approximates the joint probability between source and target languages capturing bilingual context, as described by the following equation:

$$p(s_1^J, t_1^I) = \prod_{i=1}^{K} p((s,t)_i | (s,t)_{i-N+1}, ..., (s,t)_{i-1}) \quad (1)$$

where $(s,t)_i$ refers to the i^{th} tuple of a sentence pair being segmented into K tuples. A detailed comparison between Ngram-based and phrase-based SMT can be found in [8].

2.1. Tuple extraction

Given a certain word-aligned parallel corpus, tuples are extracted according to the following constraints [9]:

- a monotonic segmentation of each bilingual sentence pair is produced
- no word in a tuple is aligned to words outside of it
- no smaller tuples can be extracted without violating the previous constraints

2.2. Feature functions

As additional feature functions to better guide the translation process, the system incorporates a target language model, a word bonus model and two lexicon models.

The *target language model* is estimated as a standard n-gram over the target words, as follows:

$$p_{LM}(t_k) \approx \prod_{n=1}^{k} p(w_n|w_{n-N+1}, ..., w_{n-1}) \tag{2}$$

where t_k refers to the partial hypothesis and w_n to the n^{th} word in it.

Usually, this feature is accompanied by a *word bonus model* based on sentence length, compensating the target language model preference for short sentences (in number of target words). This bonus depends on the number of target words in the partial hypothesis, denoted as:

$$p_{WP}(t_k) = exp(\text{number of words in } t_k) \tag{3}$$

where t_k refers to the partial hypothesis.

Finally, the third and fourth feature functions correspond to source-to-target and target-to-source *lexicon models*. These models use IBM model 1 translation probabilities to compute a lexical weight for each tuple, accounting for the statistical consistency of the pairs of words inside the tuple. These lexicon models are computed according to the following equation:

$$p_{IBM1}((s,t)_n) = \frac{1}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} p(t_n^i|s_n^j) \tag{4}$$

where s_n^j and t_n^i are the j^{th} and i^{th} words in the source and target sides of tuple $(s,t)_n$, being J and I the corresponding total number words in each side of it.

To compute the forward lexicon model, IBM model 1 lexical parameters from GIZA++ source-to-target alignments are used. In the case of the backward lexicon model, GIZA++ target-to-source alignments are used instead.

2.3. MARIE decoder

For decoding, we use MARIE [10], a freely-available tool developed at TALP Research Center, which takes all the previous models into account in an efficient beam search. For efficient pruning of the search space, *threshold pruning*, *histogram pruning* and *hypothesis recombination* are used.

Apart from monotone search, MARIE also implements full reordered search, which can be constrained by a set of parameters, as explained in the following section.

TALPtuples does not incorporate any rescoring module, therefore choosing its 1-best hypothesis as final translation solution.

3. Word ordering strategies

When dealing with pairs of languages with non-monotonic word order, a certain reordering strategy is required. Apart from that, tuples need to be extracted by an unfolding technique [11]. This means that the tuples are broken into smaller tuples, and these are sequenced in the order of the target words.

In order not to lose the information on the correct order, the decoder performs then a reordered search (or a monotone search extended with reordering paths), which is guided by the n-gram model of the unfolded tuples and the additional feature models. Figure 1 shows an example of tuple unfolding compared to the monotonic extraction. The unfolding technique produces a different bilingual n-gram language model with reordered source words.

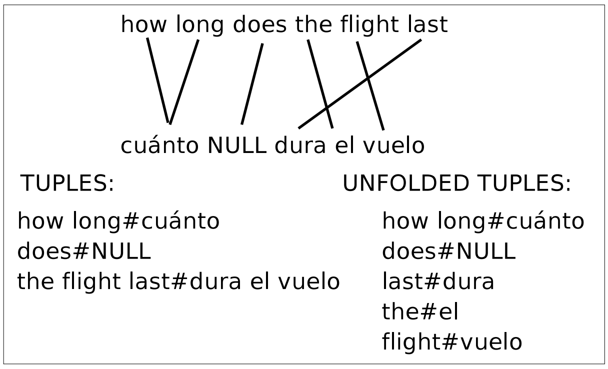

Figure 1: *Comparing regular and unfolded tuples.*

This year TALPtuples implements two reordering strategies. On the one hand, the full reordered search is constrained as done in [7]. On the other, the monotone search graph is extended by adding a few paths which reorder source words. These approaches are explained next.

3.1. Constrained reordered search

Given that the bilingual n-gram is estimated over the reordered set of tuples (unfolded tuples), a certain reordered search needs to be allowed in decoding time. However, due to the combinatory explosion if no restrictions on reordering are applied, we use two parameters to restrict the search:

- A distortion limit (m): Any source word (or tuple) is only allowed to be reordered if it does not exceed a distortion limit, measured in number of source words.
- A reordering limit (j): Any translation path is only allowed to perform j reordering jumps.

The use of these constraints implies a necessary trade-off between quality and efficiency, depending on the difficulty of the task. In our experiments for IWSLT 2006, given the average sentence length, these parameters were set to $m = 5$ and $j = 3$ for all language pairs.

3.2. Extended monotone search: reordering patterns

This reordering framework consists of using a set of automatically learnt rewrite rules to extend the monotonic search graph with reordering hypotheses (details in [12]).

A reordering pattern consists of the next rewrite rule:

$$t_1, ..., t_n \mapsto i_1, ..., i_n$$

where $t_1, ..., t_n$ is a sequence of POStags (relating a sequence of source words), and $i_1, ..., i_n$ indicates which order of the source words generate monotonically the target words.

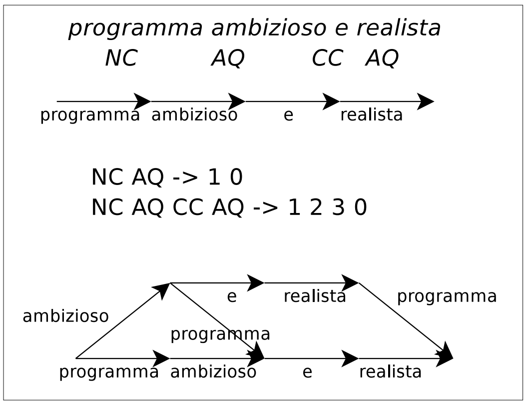

Figure 2: *Search graph extension.*

Patterns are extracted in training from the crossed links found in the word alignment, in other words, found in translation tuples (as no word within a tuple can be linked to a word out of it [9]).

Once all instances of rewrite patterns have been obtained, we compute a score for each pattern on the basis of relative frequency:

$$p(t_1, ..., t_n \mapsto i_1, ..., i_n) = \frac{N(t_1, ..., t_n \mapsto i_1, ..., i_n)}{N(t_1, ..., t_n)} \quad (5)$$

This score is used in training to prune out those patterns not achieving a threshold limit.

Starting from the monotonic graph, each sequence of input POStags fulfilling a source-side rewrite rule implies the addition of a reordering arc (which encodes the reordering detailed in the target-side of the rule). Figure 2 shows how three rewrite rules applied over an input sentence extend the search graph given the reordering patterns that match the source POS tag sequence [1].

In the search, the decoder makes use of the whole set of models to score each reordering hypothesis, mainly driven by the N-gram translation model, as it has been estimated with reordered source words.

[1]*NC*, *CC* and *AQ* stand respectively for name, conjunction and adjective.

4. Linguistic tuple segmentation

Note that the standard tuple extraction algorithm from Section 2 defines a unique set of tuples except whenever the resulting tuple contains no source word (NULL-source tuple). As these units cannot be allowed in decoding new sentences, a certain hard decision must be taken regarding tuple segmentation.

Recent results show that taking this decision depending on the forward and backward entropies of Part-Of-Speech distributions can produce a better estimated bilingual n-gram model [13]. In particular, given the tuple sequence described as follows:

$$\begin{array}{ccc} < ...s_j > & \text{NULL} & < s_{j+1}... > \\ | & | & | \\ < ...t_{i-1} > & t_i & < t_{i+1}... > \end{array}$$

where s_j means word in position j in source sentence, and equivalently t_i means word in position i in target sentence, we can define a 'forward' entropy of the POS distribution in position $i+1$ given (t_{i-1}, t_i) as in equation 6:

$$H^f_{POS} = -\sum_{POS} p^f_{POS} \log p^f_{POS} \quad (6)$$

where

$$p^f_{POS} = \frac{N(t_{i-1}, t_i, POS_{i+1})}{\sum_{POS'} N(t_{i-1}, t_i, POS'_{i+1})} \quad (7)$$

is the probability of observing a certain Part-Of-Speech *following* the sequence of words defined by t_i and t_{i+1}.

Equivalently, we can define a 'backward' entropy of the POS distribution in position $i-1$ given (t_i, t_{i+1}) as in equation 8:

$$H^b_{POS} = -\sum_{POS} p^b_{POS} \log p^b_{POS} \quad (8)$$

where

$$p^b_{POS} = \frac{N(POS_{i-1}, t_i, t_{i+1})}{\sum_{POS'} N(POS'_{i-1}, t_i, t_{i+1})} \quad (9)$$

is the probability of observing a certain Part-Of-Speech *preceding* the sequence of words defined by t_{i-1} and t_i.

Then, we can take a tuple segmentation decision favoring the most POS-entropic case. The rationale behind this is that, if $H^f_{POS} > H^b_{POS}$, we have observed the first sequence of words comprised of (t_{i-1}, t_i) in more grammatically different situations than the latter sequence comprised of (t_i, t_{i+1}). Therefore, we can induce that t_{i-1} and t_i tend to be more often connected than t_i and t_{i+1}, and should belong to the same translation tuple. Analogously, one can conclude the contrary if $H^f_{POS} < H^b_{POS}$.

5. Experiments

In this section all the experimental work conducted for IWSLT 2006 shared tasks is reported. TALPtuples participated in the open-data track for all translation directions (from Arabic, Chinese, Italian and Japanese into English). In all four cases, the 1-best speech recognition output was taken as input to the translation system. Therefore, no n-best list nor word graph were used.

5.1. Tasks description

For internal development work, true case and punctuation marks were removed from all parallel corpora (train, develop, test and references), therefore optimizing according to the 'additional' scoring scheme as defined in IWSLT 2006, consistent with previous years. For the final evaluation test set, punctuation marks and true case were included by using SRILM 'disambig' tool as suggested by IWSLT organizers.

Given the availability of up to four development sets for all language pairs, our strategy was to use development 4 as internal development set (**dev4**), while randomly selecting 500 sentences from development 1, 2 and 3 (around 160 sentences from each) to build an internal test set (**dev123**). Finally, the roughly 1k remaining development sentences were included in the training corpus by selecting the first English manual reference.

		sent.	wrds	voc.	slen.	refs.
train	ar	24.0k	183k	10.5k	7.6	1
	en		166k	7.3k	6.9	
dev4	ar	489	5,889	1,237	12	7
dev123	ar	500	3,329	1,037	6.7	16
test	ar	500	6,570	1,480	13.1	7
ASRtest	ar	500	6,659	1,532	13.3	7

Table 1: *Arabic→English corpus statistics.*

In all experiments the union alignment was used to extract unfolded tuples. Following the consensus strategy proposed in [14], optimization criterion was set to $100 \cdot BLEU + 4 \cdot NIST$ on the **dev4** set. These parameters were equally used for translating correctly recognized text our 1-best recognition results.

		sent.	wrds	voc.	slen.	refs.
train	zh	46.9k	314k	9.7k	6.7	1
	en		326k	9.6k	7.0	
dev4	zh	489	5,478	1,096	11.2	7
dev123	zh	500	3,005	909	6.0	16
test	zh	500	5,846	1,292	11.7	7
ASRtest	zh	500	5,825	1,311	11.6	7

Table 2: *Chinese→English corpus statistics.*

Corpora statistics for all language pairs can be found in Tables 1, 2, 3 and 4, respectively, where number of sentences, running words, vocabulary, sentence length and human references are shown.

		sent.	wrds	voc.	slen.	refs.
train	it	24.6k	155k	10.2k	6.3	1
	en		166k	7.3k	6.8	
dev4	it	489	5,193	1,192	10.6	7
dev123	it	500	2,807	969	5.6	16
test	it	500	5,978	1,429	12.0	7
ASRtest	it	500	5,767	1,517	11.5	7

Table 3: *Italian→English corpus statistics.*

		sent.	wrds	voc.	slen.	refs.
train	jp	45.2k	390k	10.6k	8.6	1
	en		325k	9.6k	7.2	
dev4	jp	489	6,758	1,169	13.8	7
dev123	jp	500	3,818	936	7.6	16
test	jp	500	7,367	1,301	14.7	7
ASRtest	jp	500	7,494	1,331	15.0	7

Table 4: *Japanese→English corpus statistics.*

5.2. Language-dependent preprocessing

For all language pairs, training sentences were split by using final dots on both sides of the bilingual text (when the number of dots was equal), increasing the number of sentences and reducing its length. Specific preprocessing for each language is detailed in the following respective section.

5.2.1. English

English preprocessing includes Part-Of-Speech tagging using freely-available *TnT* tagger [15] and lemmatization using *wnmorph*, included in the WordNet package [16]. The English Penn Treebank Tag Set used contains 36 different tags.

5.2.2. Arabic

Following a similar approach to that in [17], we use the Buckwalter Arabic Morphological Analyzer[2] to obtain possible word analyses for Arabic, and disambiguate them using the Morphological Analysis and Disambiguation for Arabic (MADA) tool [18], kindly provided by the University of Columbia.

Once analyzed, Arabic words are segmented by separating all prefixes (prepositions, conjunctions, the article and the future marker) and suffixes (pronominal clitics). The tool also provides POS tags for the resultant tokens. The Arabic Treebank tag set used contains 20 different tags.

[2]Version 2.0. Linguistic Data Consortium Catalog: LDC2004L02.

5.2.3. Chinese

Chinese preprocessing included resegmentation and POS-tagging. These tasks were done by using ICTCLAS [19]. Resultant tag set has a vocabulary of 41 different tags.

5.2.4. Italian

Italian has been POS-tagged and lemmatized using the freely-available FreeLing morpho-syntactic analysis package [20]. As tags contain rich morphological features such as person, gender or number, tag set contains 272 different tags.

Additionally, Italian contracted prepositions have been separated into preposition + article, such as 'alla'→'a la', 'degli'→'di gli' or 'dallo'→'da lo', among others.

5.2.5. Japanese

Japanese language is a specific task for SMT due to absence of delimiters between words. We addressed this issue by word segmentation using the freely available JUMAN tool [21] version 5.1. This tool was also used for POS-tagging of the Japanese text, using a tag set of 15 different tags.

official	test		ASRtest	
	BLEU	NIST	BLEU	NIST
Arabic→English				
P: m5j3	**0.232**	**6.24**	**0.214**	**5.82**
C1: rgraph	0.227	6.14	0.205	5.69
C2: m5j3 segIBM	0.227	6.06	0.210	5.63
C3: m5j3 lm20	0.225	6.13	0.205	5.71
Chinese→English				
P: m5j3	**0.186**	5.57	**0.162**	4.98
C1: rgraph	0.183	**5.74**	0.157	**5.12**
Italian→English				
P: rgraph alem	**0.333**	**7.75**	**0.282**	**6.87**
C1: rgraph	0.331	7.63	0.278	6.75
C2: rgraph segIBM	0.332	7.64	0.273	6.71
C3: rgraph lm20	0.323	7.54	0.271	6.69
Japanese→English				
P: rgraph	0.146	**5.27**	0.137	**4.94**
C1: m5j3	**0.152**	5.18	**0.141**	4.89

Table 5: *Translation results for IWSLT 2006 tasks.*

5.3. Results

TALPtuples official evaluation results for correct text (**test**) and 1-best speech recognition output (**ASRtest**) are shown in Table 5. For each language pair, the submitted runs correspond to those experiments obtaining best performance in internal development and test sets (**dev123** and **dev4**). 'P' denotes the primary run and 'C_x' denote the contrastive runs.

By default, tuples are assumed to be generated from the union alignment, except when 'alem' is stated, indicating that the alignment was performed on the lemmas instead of the words, following the approach presented in [22]. Apart from that, tuple segmentation using entropy of the POS distribution (described in Section 4) is always assumed, unless otherwise stated ('segIBM' refers to tuple segmentation using the IBM model 1 criterion from [7]).

When it comes to training the target language model, the ∼40k English sentences from the Chinese–English parallel corpus are always used, even for those pairs of languages where parallel text is smaller (Arabic and Italian). For these languages, an additional experiment training the target language model only with the English sentences of its parallel text is also shown (marked with 'lm20').

Finally, the two reordering strategies presented in Section 3 are denoted as 'm5j3' for constrained reordered search (referring to the constraining parameters) and 'rgraph' for the use of reordering patterns to extend the monotone search graph.

5.4. Discussion

In the Arabic→English task, the 'm5j3' reordering strategy (P experiment) seems to perform best. Although differences with 'rgraph' (C1) are small, we can conclude that the proposed POS reordering patterns do not capture enough of the reordering needs for this language pair. Regarding tuple segmentation, the novel linguistic tuple segmentation outperforms previous IBM-based approach (C2). Finally, the use of the extended target language model proves useful to achieve significant improvement (when comparing P to C3), probably because it belongs to the same domain.

In Chinese→English and Japanese→English, both reordering strategies behave similarly in terms of translation quality. Whereas 'm5j3' tends to produce slightly higher BLEU scores, 'rgraphs' obtained slightly better NIST scores. Although further research is needed to improve on this difficult subject, these results are very positive, especially when taking efficiency into account. In fact, the extended monotone search with reordering patterns is much more efficient than reordered search, as also studied in [12].

Finally, in the Italian→English task, the 'rgraph' strategy significantly outperformed all experiments with reordered search in internal development work, so no experiment with 'm5j3' configuration was submitted. This correlates with the fact that this approach also works very well for a Spanish→English task [12], and that Spanish and Italian exhibit a big structural similarity.

In this case, the alignment based on lemmas instead of words did produce a slight improvement in performance (comparing P and C1). Again, POS-based tuple segmentation was slightly better than IBM-based (comparing C1 and C2), except for the correct recognition test, where it basically performed the same. And finally, the contribution of the extended target language model yielded again significant improvement, as observed when comparing C1 and C3.

6. Conclusions and further work

In this paper we introduced **TALPtuples**, the Ngram-based SMT system of the TALP Research Center (UPC, Barcelona) participating in IWSLT 2006. Apart from briefly summarizing the system architecture from the previous year evaluation, special emphasis was put on describing the novel features of the system.

These were basically two; on the one hand, a novel reordering strategy based on extending the monotone search graph with a few relevant reordered paths (which are automatically learnt from word alignment and source-language POS sequences), and on the other hand, an advanced tuple segmentation criterion based on entropy of POS distributions.

Main conclusions from the official results for the four language pairs are:

- Regarding reordering, the novel approach by incorporating reordering patterns outperforms reordered search for Italian→English, achieves similar results for Chinese→English and Japanese→English and is slightly worse in Arabic→English. For those languages pairs demanding long reorderings, the pattern definition as sequences of Part-Of-Speech tags seems to be leading to sparseness. Further research should therefore focus on pattern extraction for these language pairs

- Regarding tuple segmentation, the novel tuple segmentation based on POS entropy yields a slight yet systematic improvement in translation quality

Additionally, further research should be directed towards more integration of speech recognition output and the SMT system, as neither word lattices nor N-best lists were used as input to the translation module for any of these experiments. This would probably lead to improved performance.

7. Acknowledgements

This work was partially funded by the European Union under the integrated project TC-STAR (IST-2002-FP6-506738, `http://www.tc-star.org`), by Generalitat de Catalunya, by the European Social Fund and by Universitat Politècnica de Catalunya.

The authors want to explicitly thank Nizar Habash for his readiness to make MADA available for Arabic experiments.

8. References

[1] E. Vidal, "Finite-state speech-to-speech translation," *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 111–114, April 1997.

[2] A. de Gispert and J. Mariño, "Using X-grams for speech-to-speech translation," *Proc. of the 7th Int. Conf. on Spoken Language Processing, ICSLP'02*, pp. 1885–1888, September 2002.

[3] M. Eck and C. Hori, "Overview of the IWSLT 2005 Evaluation Campaign," *Proc. of the 2nd Int. Workshop on Spoken Language Translation, IWSLT'05*, pp. 11–32, October 2005.

[4] P. Koehn and C. Monz, "Shared task: Statistical Machine Translation between European Languages," *Proc. of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pp. 119–124, June 2005.

[5] ——, "Manual and automatic evaluation of machine translation between european languages," in *Proceedings of the Workshop on Statistical Machine Translation*, New York City, June 2006, pp. 102–21.

[6] J. Mariño, R. Banchs, J. Crego, A. de Gispert, P. Lambert, M. Costa-jussà, and J. Fonollosa, "Bilingual N–gram statistical machine translation," *Proc. of the MT Summit X*, pp. 275–282, September 2005.

[7] J. Crego, A. de Gispert, and J. Mariño, "TALP: The UPC tuple-based SMT system," *Proc. of the 2nd Int. Workshop on Spoken Language Translation, IWSLT'05*, pp. 191–198, October 2005.

[8] J. Crego, M. Costa-jussà, J. Mariño, and J. Fonollosa, "Ngram-based versus phrase-based statistical machine translation," *Proc. of the 2nd Int. Workshop on Spoken Language Translation, IWSLT'05*, pp. 177–184, October 2005.

[9] J. Crego, J. Mariño, and A. de Gispert, "Finite-state-based and phrase-based statistical machine translation," *Proc. of the 8th Int. Conf. on Spoken Language Processing, ICSLP'04*, pp. 37–40, October 2004.

[10] ——, "An ngram-based statistical machine translation decoder," *Proc. of the 9th European Conference on Speech Communication and Technology, Interspeech'05*, pp. 3193–3196, September 2005.

[11] ——, "Reordered search and tuple unfolding for ngram-based smt," *Proc. of the MT Summit X*, pp. 283–89, September 2005.

[12] J. Crego and J. Mariño, "Reordering experiments for N-gram-based SMT," *1st IEEE/ACL Workshop on Spoken Language Technology*, December 2006.

[13] A. de Gispert and J. Mariño, "Linguistic tuple segmentation in ngram-based statistical machine translation," *Proc. of the 9th Int. Conf. on Spoken Language Processing, ICSLP'06*, pp. 1149–1152, September 2006.

[14] B. Chen, R. Cattoni, N. Bertoldi, M. Cettolo, and M. Federico, "The ITC-irst SMT system for IWSLT-2005," *Proc. of the 2nd Int. Workshop on Spoken Language Translation, IWSLT'05*, pp. 98–104, October 2005.

[15] T. Brants, "TnT – a statistical part-of-speech tagger," in *Proc. of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, WA, 2000.

[16] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, and R. Tengi, "Five papers on Wordnet," *Special Issue of International Journal of Lexicography*, vol. 3, no. 4, pp. 235–312, 1991.

[17] N. Habash and F. Sadat, "Arabic preprocessing schemes for statistical machine translation," in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, New York City, USA, June 2006, pp. 49–52.

[18] N. Habash and O. Rambow, "Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop," in *43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, MI, June 2005, pp. 573–580.

[19] H. Zhang, H. Yu, D. Xiong, and Q. Liu, "HHMM-based chinese lexical analyzer ICTCLAS," in *Proc. of the 2nd SIGHAN Workshop on Chinese language processing*, Sapporo, Japan, 2003, pp. 184–187.

[20] J. Atserias, B. Casas, E. Comelles, M. González, L. Padró, and M. Padró, "Freeling 1.3: Syntactic and semantic services in an open-source NLP library," *5th Int. Conf. on Language Resources and Evaluation, LREC'06*, pp. 2281–2286, May 2006.

[21] Y. Matsumoto and M. Nagao, "Improvements of Japanese morphological analyzer JUMAN," in *Proc. of the Int. Workshop on Sharable Natural Language Resources*, August 1994, pp. 22–28.

[22] A. de Gispert, D. Gupta, M. Popović, P. Lambert, J. Mariño, M. Federico, H. Ney, and R. Banchs, "Improving statistical word alignments with morpho-syntactic transformations," *Proceedings of 5th International Conference on Natural Language Processing, FinTAL'06*, pp. 368–379, August 2006.