# The TALP-UPC phrase-based translation systems for WMT12: Morphology simplification and domain adaptation

**Lluís Formiga, Carlos A. Henríquez Q., Adolfo Hernández,**
**José B. Mariño, Enric Monte and José A. R. Fonollosa**
TALP Research Centre
Universitat Politècnica de Catalunya
Barcelona, Spain
{lluis.formiga,carlos.henriquez,adolfo.hernandez
jose.marino,enric.monte,jose.fonollosa}@upc.edu

## Abstract

This paper describes the UPC participation in the WMT 12 evaluation campaign. All systems presented are based on standard phrase-based Moses systems. Variations adopted several improvement techniques such as morphology simplification and generation and domain adaptation. The morphology simplification overcomes the data sparsity problem when translating into morphologically-rich languages such as Spanish by translating first to a morphology-simplified language and secondly leave the morphology generation to an independent classification task. The domain adaptation approach improves the SMT system by adding new translation units learned from MT-output and reference alignment. Results depict an improvement on TER, METEOR, NIST and BLEU scores compared to our baseline system, obtaining on the official test set more benefits from the domain adaptation approach than from the morphological generalization method.

## 1 Introduction

TALP-UPC (Center of Speech and Language Applications and Technology at the Universitat Politècnica de Catalunya) has participated in the WMT12 shared task translating across two directions: English to Spanish and Spanish to English tasks.

For the Spanish to English task we submitted a baseline system that uses all parallel training data and a combination of different target language models (LM) and Part-Of-Speech (POS) language models. A similar configuration was submitted for the English to Spanish task as baseline. Our main approaches enriched the latter baseline in two independent ways: morphology simplification and domain adaptation by deriving new units into the phrase-table. Furthermore, additional specific strategies have been addressed on all systems to deal with well known linguistic phenomena in Spanish such as clitics and contractions.

The paper is presented as follows. Section 2 presents the main rationale for the phrase-based system and the main pipeline of our baseline system. Section 3 presents the approaches taken to improve the baseline system on the English to Spanish task. Section 4 presents the obtained results on internal and official test sets while conclusions and further work are presented in Section 5.

## 2 Baseline system: Phrase-Based SMT

Classically, a phrase-based translation system implements a log-linear model in which a foreign language sentence $f_1^j = f_1, f_2, \ldots, f_j$ is translated into another language sentence $e_1^I = e_1, e_2, \ldots, e_I$ by searching for the translation hypothesis that maximizes a log-linear combination of feature models (Brown et al., 1990):

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

where the separate feature functions $h_m$ refer to the system models and the set of $\lambda_m$ refers to the weights corresponding to these models. As feature functions we used the standard models available

*Proceedings of the 7th Workshop on Statistical Machine Translation*, pages 275–282,
Montréal, Canada, June 7-8, 2012. ©2012 Association for Computational Linguistics

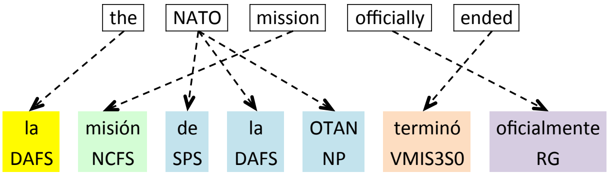

Figure 1: Factored phrase-based MT based on translation from surface to surface and Part-of-Speech

on Moses, i.e., relative frequencies, lexical weights, word and phrase penalty, *wbe-msd-bidirectional-fe* reordering models and two language models, one for surface and one for POS tags. Phrase scoring was computed using Good-Turing discounting (Foster et al., 2006).

The tuning process was done using MERT (Och, 2003) with Minimum Bayes-Risk decoding (MBR) (Kumar and Bryne, 2004) on Moses and focusing on minimizing the BLEU score (Papineni et al., 2002) of the development set. Final translations were also computed using MBR decoding.

Additionally to the settings mentioned before, we worked with a factored version of the corpus. Factored corpora augments surface forms with additional information, such as POS tags or lemmas as shown in Figure 1. In that case, factors other than surface (e.g. POS) are usually less sparse, allowing to build factor-specific language models with higher-order n-grams. These higher-order language models usually help to obtain more syntactically correct output. Concretely we map input source surfaces to target surfaces and POS tags.

### 2.1 Corpus used

The baseline system was trained using all parallel corpora, i.e. the European Parliament (EPPS) (Koehn, 2005), News Commentary and United Nations. Table 1 shows the statistics of the training data after the cleaning process described later on Subsection 2.2.

Regarding the monolingual data, there was also more News corpora separated by years for Spanish and English and there was the Gigaword monolingual corpus for English. All data can be found on the Translation Task's website[1]. We used all News corpora (and Gigaword for English) to build the language model. Initially, a LM was built for every corpus and then they were combined to produce de final LM. Table 2 presents the statistics of each corpora, again after the cleaning process.

| Corpus | | Sent. | Words | Vocab. | avg.len. |
|---|---|---|---|---|---|
| EPPS | Eng | 1.90 M | 49.40 M | 124.03 k | 26.05 |
| | Spa | | 52.66 M | 154.67 k | 27.28 |
| News.Com | Eng | 0.15 M | 3.73 M | 62.70 k | 24.20 |
| | Spa | | 4.33 M | 73.97 k | 28.09 |
| UN | Eng | 8.38 M | 205.68 M | 575.04 k | 24.54 |
| | Spa | | 239.40 M | 598.54 k | 28.56 |

Table 1: English-Spanish corpora statistics for NAACL-WMT 2012 after cleaning process

| Corpus | | Sent. | Words | Vocab. |
|---|---|---|---|---|
| EPPS | Eng | 2.22 M | 59.88 M | 144.03 k |
| | Spa | 2.12 M | 61.97 M | 174.92 k |
| News.Com. | Eng | 0.21 M | 5.08 M | 72.55 k |
| | Spa | 0.18 M | 5.24 M | 81.56 k |
| UN | Eng | 11.20 M | 315.90 M | 767.12 k |
| | Spa | 11.20 M | 372.21 M | 725.73 k |
| News.07 | Eng | 3.79 M | 90.25 M | 711.55 k |
| | Spa | 0.05 M | 1.33 M | 64.10 k |
| News.08 | Eng | 13.01 M | 308.82 M | 1555.53 k |
| | Spa | 1.71 M | 49.97 M | 377.56 k |
| News.09 | Eng | 14.75 M | 348.24 M | 1648.05 k |
| | Spa | 1.07 M | 30.57 M | 287.81 k |
| News.10 | Eng | 6.81 M | 158.15 M | 915.14 k |
| | Spa | 0.69 M | 19.58 M | 226.76 k |
| News.11 | Eng | 13.46 M | 312.50 M | 1345.79 k |
| | Spa | 5.11 M | 151.06 M | 668.63 k |
| Giga | Eng | 22.52 M | 657.88 M | 3860.67 k |

Table 2: Details of monolingual corpora used for building language-models.

For internal testing we used the News 2011's data and concatenated the remaining three years of News data as a single parallel corpus for development. Table 3 shows the statistics for these two sets and includes in the last rows the statistics of the official test set for this year's translation task.

### 2.2 Corpus processing

All corpora were processed in order to remove or normalize ambiguous or special characters such as quotes and spaces. Among other TALP-UPC specific scripts, we used a modified version of the normalized-punctuation script provided by the organizers in order to skip the reordering rules which involved quotes and stop punctuation signs.

[1] http://www.statmt.org/wmt12/translation-task.html

| Corpus | | Sent. | Words | Vocab. | avg.len. |
|---|---|---|---|---|---|
| dev | Eng | 7.57 k | 189.01 k | 18.61 k | 24.98 |
| | Spa | | 202.80 k | 21.75 k | 26.80 |
| test11 | Eng | 3.00 k | 74.73 k | 10.82 k | 24.88 |
| | Spa | | 81.01 k | 12.16 k | 26.98 |
| test12 | Eng | 3.00 k | 72.91 k | 10.24 k | 24.28 |
| | Spa | | 80.38 k | 12.02 k | 26.77 |

Table 3: Detail of development and test corpora used to tune and test the system.

POS-Tagging and tokenization for both Spanish and English data sets were obtained using FreeLing (Padró et al., 2010). Freeling tokenization is able to deal with contractions (“del” → “de el”) and clitics separation (“cómpramelo” → “compra me lo”) in Spanish and English. Stemming was performed using Snowball (Porter, 2001).

Surface text was lowercased conditionally based on the POS tagging: proper nouns and adjectives were separated from other POS categories to determine if a string should be fully lowercased (no special property), partially lowercased (proper noun or adjective) or not lowercased at all (acronym).

Bilingual corpora were cleaned with *clean-corpus-n* script of Moses (Koehn et al., 2007) removing all sentence pair with more than 70 words in any language, considering the already tokenized data. That script also ensures a maximum length ratio below of nine (9) words between source and target sentences.

Postprocessing in both languages consisted of a recasing step using Moses recaser script. Furthermore we built an additional script in order to check the casing of output names with respect to source sentence names and case them accordingly, with exception of names placed at beginning of the sentence. After recasing, a final detokenization step was performed using standard Moses tools. Spanish postprocessing also included two special scripts to recover contractions and clitics.

### 2.3 Language Model and alignment configuration

Word alignment was performed at stem level with GIZA++ toolkit (Och and Ney, 2003) and *grow-diag-final-and* joint alignment.

Language models were built from the monolingual data provided covering different domains: Europarl, News and UN. We built them using Kneser-Ney algorithm (Chen and Goodman, 1999), interpolation in order to avoid over-fitting and considering unknown words. First we built a 5-gram language model for each corpus; then, the final LM was obtained interpolating them all towards the development set. We used SRI Language Model (Stolcke, 2002) toolkit, which provides *compute-best-mix* script for the interpolation.

The POS language model was built analogously to the surface language with some variants: it was a 7-gram LM, without discounting nor interpolation.

## 3 Improvement strategies

### 3.1 Motivations

In order to improve the baseline system we present two different strategies. First we present an improvement strategy based on morphology simplification plus generation to deal with the problems raised by morphological rich languages such as Spanish. Second we present a domain adaptation strategy that consists in deriving new units into the phrase-table.

### 3.2 Morphology simplification

The first improvement strategy is based on morphology simplification when translating from English to Spanish.

The problems raised when translating from a language such as English into richer morphology languages are well known and are a research line of interest nowadays (Popovic and Ney, 2004; Koehn and Hoang, 2007; de Gispert and Mariño, 2008; Toutanova et al., 2008; Clifton and Sarkar, 2011). In that direction, inflection causes a very large target-language lexicon with a significant data sparsity problem. In addition, system output is limited only to the inflected phrases available in the parallel training corpus. Hence, SMT systems cannot generate proper inflections unless they have learned them from the appropriate phrases. That would require to have a parallel corpus containing all possible word inflections for all phrases available, which it is an unfeasible task.

The morphology related problems in MT have been addressed from different approaches and may

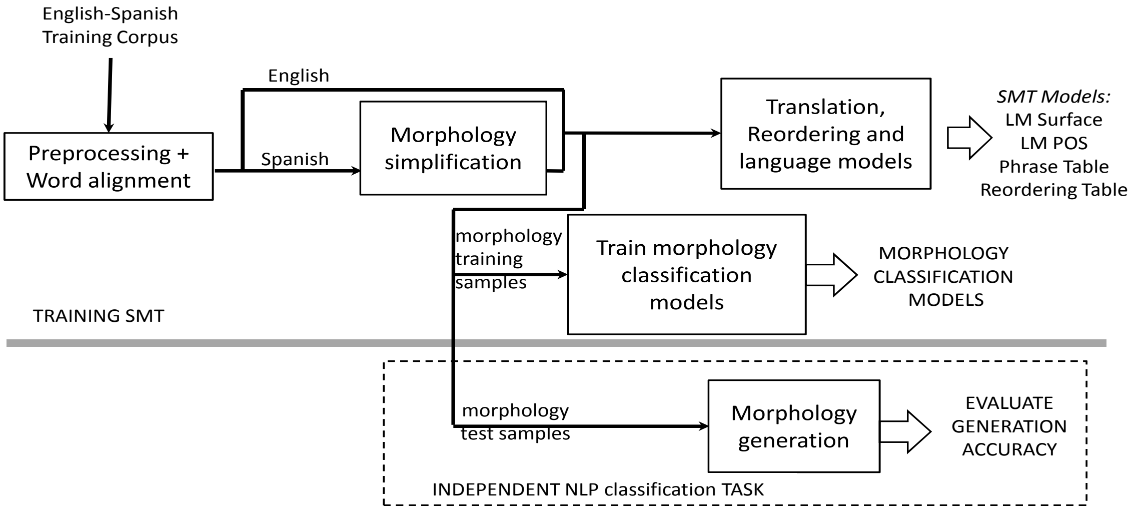

Figure 2: Above, flow diagram of the training of simplified morphology translation models. Below, Spanish morphology generation as an independent classification task.

| Type | Text |
|---|---|
| *PLAIN TARGET:* | la Comisión **puede** llegar a paralizar el programa |
| *TARGET+PoS (Gen. Sur.):* | la Comisión **VMIP3S0[poder]** llegar a paralizar el programa |
| *TARGET+PoS (Simpl. PoS):* | la Comisión **VMIPpn0[poder]** llegar a paralizar el programa |

Table 4: Example of morphology simplification steps taken for Spanish verbs.

be summarized in four categories: *i*) factored models (Koehn and Hoang, 2007), enriched input models (Avramidis and Koehn, 2008; Ueffing and Ney, 2003), segmented translation (Virpioja et al., 2007) and morphology generation (Toutanova et al., 2008; de Gispert and Mariño, 2008).

Our strategy for dealing with morphology generation is based in the latter approach (de Gispert and Mariño, 2008) (Figure 2). We center our strategy in simplifying only verb forms as previous studies indicate that they contribute to the main improvement (Ueffing and Ney, 2003; de Gispert and Mariño, 2008). That strategy makes clear the real impact of morphology simplification by providing an upper bound oracle for the studied scenarios.

The approach is as follows: First, target verbs are simplified substituting them with their simplified forms (Table 4). In this example, the verb form 'puede' (he can) is transformed into 'VMIPpn0[poder]', indicating simplified POS and base form (lemma); where 'p' and 'n' represent any person and number once simplified (from 3rd person singular). Secondly, standard MT models are obtained from English into simplified morphology Spanish. Morphology prediction acts as a black box, with its models estimated over a simplified morphology parallel texts (including target language model and lexicon models).

Generation is implemented by Decision Directed Acyclic Graphs (DDAG) (Platt et al., 2000) compound of binary SVM classifiers. In detail, a DDAG combines many two-class classifiers to a multi-classification task (Hernández et al., 2010).

### 3.3 Domain adaptation

Depending on the available resources, different domain adaptation techniques are possible. Usually, the baseline system is built with a large out-of-domain corpus (in our case the European Parliament) and we aim to adapt to another domain that has limited data, either only monolingual or hopefully bilingual as well. The WMT Translation Task focuses on adapting the system to a news domain, offering an in-domain parallel corpus to work with.

In case of additional target monolingual data, previous works have focused on language model interpolations (Bulyko et al., 2007; Mohit et al., 2009; Wu et al., 2008). When parallel in-domain data is available, the latest researches have focused on mixture model adaptation of the translation model (Civera and Juan, 2007; Foster and Kuhn, 2007; Foster et al., 2010). Our work is closer to the latest approaches. We used the in-domain parallel data to adapt the translation model, but focusing on the decoding errors that the out-of-domain baseline system made while translating the in-domain corpus. The idea is to detect where the system made its mistakes and use the in-domain data to teach it how to correct them.

Our approach began with a baseline system built with the Parliament and the United Nations parallel corpora but without the News parallel corpus. The rest of the configuration remained the same for the baseline. With this alternative baseline system, we translated the source side of the News parallel corpus to obtain a revised corpus of it, as defined in (Henríquez Q. et al., 2011). The revised corpus consists of the source side, the output translation and the target side, also called the target correction. The output translation and its reference are then compare to detect possible mistakes that the system caused during decoding.

The translation was used as a pivot to find a word-to-word alignment between the source side and the target correction. The word-to-word alignment between source side and translation was provided by Moses during decoding. The word-to-word alignment between the output translation and target correction was obtained following these steps:

1. Translation Edit Rate (Snover et al., 2006) between each output translation and target correction sentence pair was computed to obtain its edit path and detect which words do not change between sentences. Words that did not change were directly linked

2. Going from left to right, for each unaligned word $w_{out}$ on the output translation sentence and each word $w_{trg}$ on the target correction sentence, a similarity function was computed between them and $w_{out}$ got aligned with the word $w_{trg}$ that maximized this similarity.

The similarity function was defined as a linear combination of features that considered if the words $w_{out}$ and $w_{trg}$ were identical, if the previous or following word of any of them were aligned with each other and a lexical weight between them using the bilingual lexical features from the baseline as references.

With both word-to-word alignments computed for a sentence pair, we linked source word $w_{src}$ with target word $w_{trg}$ is and only if exists a output translation word $w_{out}$ such that there is a link between $w_{src}$ and $w_{out}$ and a link between $w_{out}$ and $w_{trg}$.

After aligning the corpus, we built the translation and reordering model of it, using the baseline settings. We called these translation and reordering models, revised models. They include phrases found in the baseline that were correctly chosen during decoding and also new phrases that came from the differences between the output translation and its correction.

Finally, the revised translation model features were linearly combined with their corresponding baseline features to build the final translation model, called the derived translation model. The combination was computed in the following way:

$$h_d^i(s,t) = \alpha h_b^i(s,t) + (1-\alpha) h_r^i(s,t) \tag{2}$$

where $h_d^i(s,t)$ is the derived feature function $i$ for the bilingual phrase $(s,t)$, $h_b^i(s,t)$ is the baseline feature function of and $h_r^i(s,t)$ the revised feature function. A value of $\alpha = 0.60$ was chosen after determining it was the one that maximized the BLEU score of the development set during tuning. Different values for $\alpha$ were considered, between $0.50$ and $0.95$ with increments of $0.05$ between them.

Regarding the reordering model, we added the unseen phrases from the revised reordering model into the baseline reordering model, leaving the remaining baseline phrase reordering weights intact.

## 4 Results

### 4.1 Language Model perplexities

| LM | Perplexity | |
|---|---|---|
| | Surface | POS |
| Baseline | 205.36 | 13.23 |
| Simplified | 193.66 | 12.66 |

Table 6: Perplexities obtained across baseline and morphology simplification.

Before evaluating translation performance, we studied to what extent the morphology simplifica-

| EN→ES | | BLEU | | NIST | | TER | METEOR |
|---|---|---|---|---|---|---|---|
| | | CS | CI | CS | CI | CS | CI |
| test11 | Baseline | 30.7 | 32.53 | 7.820 | 8.120 | 57.19 | 55.05 |
| | Morph. Oracle | 31.56 | 33.35 | 7.949 | 8.233 | 56.44 | – |
| | Morph. Gen. | 31.03 | 32.85 | 7.866 | 8.163 | 56.95 | 55.39 |
| | Adaptation | 31.16 | 32.93 | 7.857 | 8.155 | 56.88 | 55.19 |
| test12 | Baseline | 31.21 | 32.74 | 7.981 | 8.244 | 55.76 | 55.48 |
| | Morph. Oracle | 32 | 33.41 | 8.090 | 8.339 | 55.15 | – |
| | Morph. Gen. | 31.46 | 32.98 | 8.010 | 8.274 | 55.62 | 55.66 |
| | Adaptation | 31.73 | 33.24 | 8.037 | 8.294 | 55.37 | 55.82 |

(a) English→Spanish

| ES→EN | | BLEU | | NIST | | TER | METEOR |
|---|---|---|---|---|---|---|---|
| | | CS | CI | CS | CI | CS | CI |
| test11 | Baseline | 28.81 | 30.29 | 7.670 | 7.933 | 59.01 | 51.09 |
| test12 | | 32.27 | 33.81 | 8.014 | 8.282 | 56.26 | 53.96 |

(b) Spanish→English

Table 5: Automatic scores for English↔Spanish translations. CS and CI indicate Case-Sensitive or Case-Insensitive evaluations.

tion strategy may help decreasing the language models perplexity.

In table 6 we can see the effects of simplification. Perplexity is computed from the corresponding internal test sets to the baseline or simplified language models.

In general terms, the simplification process is slightly effective, yielding an averaged improvement of $-5.02\%$.

### 4.2 Translation performance

Evaluations were performed with different translation quality measures: BLEU, NIST, TER and METEOR (Denkowski and Lavie, 2011) which evaluate distinct aspects of the quality of the translations. First we evaluated the WMT11 test (test11) as an internal indicator of our systems. Later we did the same analysis with the WMT12 official test files.

Table 5 presents the obtained results. Experiments began building the baseline system, which included the special treatment for clitics, contractions and casing as described in Section 2.2. Once the baseline was set, we proceeded with two parallel lines, one for morphology simplification and the other for domain adaptation.

For morphology generation approach (Table 5) oracles (Morph. Oracle) represent how much gain we could expect with a perfect generation module and generation (Morph. Gen.) represent the actual performance combining simplification and the generation strategies. Oracles achieve a promising averaged improvement of +1.79% (depending on the metric or the test set) with respect to the baseline. However, generation only improves the baseline by a +0.61%, encouraging us to keep working on that strategy.

Regarding the domain adaptation approach, we evaluated the internal test set (test11). As we can see again on Table 5a the adaptation strategy outperforms the baseline on all quality measures starting with an averaged gain of $+0.94\%$.

Comparing the two approaches, we can see that the domain adaptation method was better in terms of BLEU score and TER than the morphology generation but the latter was better on NIST and METEOR on our internal test set. This made us decided for the latter as the primary system submitted, leaving the domain adaptation approach system as a contrastive submission. Additionally to the automatic quality measures, we are particularly interested in the manual evaluation results, as we believe the morphology generation will be more sensitive to this type of evaluation than to automatic metrics.

Official results (test12) can be found on Table 5b. Surprisingly, this time the domain adaptation approach performed better than the morphology simplification on all metrics: BLEU, NIST, TER and METEOR, with an averaged gain of $+1.04\%$ over the baseline system, which ranks our submissions second and third in terms of BLEU scores (contrastive and primary respectively) when compared with all other submissions for the WMT12 translation task.

## 5 Conclusions and further work

This papers describes the UPC participation during the 2012 WMT's Translation Task. We have participated with a baseline system for Spanish-to-English, a baseline system for English-to-Spanish and two independent enhancements to the baseline system for English-to-Spanish as well.

Our primary submission applied morphology simplification and generation with the objective of ease the translation process when dealing with rich morphology languages like Spanish, deferring the morphology generation as an external post-process classification task.

The second approach focused on domain adaptation. Instead of concatenating the training News parallel data together with the European Parliament and United Nations, a preliminary system was built with the latter two and separated translation and reordering models were computed using the News parallel data. These models were then added to the preliminary models in order to build the adapted system.

Results showed that both approaches performed better than the baseline system, being the domain adaptation configuration the one that performed better for 2012 test in terms of all automatic quality indicators: BLEU, NIST, TER and METEOR. We look forward the the manual evaluation results as we believe our primary system may be more sensitive to this type of human evaluation.

Future work should focus on combining the two approaches, applying first morphological generalization to the training data and then using the domain adaptation technique on the resulting corpora in order to determine the joined benefits of both strategies.

## Acknowledgments

This work has been partially funded by the Spanish Government (Buceador, TEC2009-14094-C04-01) and the European Community's FP7 program under grant agreement number 247762 (FAUST, FP7-ICT-2009-4-247762).

## References

E. Avramidis and P. Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. *Proceedings of ACL-08: HLT*, pages 763–770.

P.F. Brown, J. Cocke, S.A.D. Pietra, V.J.D. Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, and P.S. Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Ivan Bulyko, Spyros Matsoukas, Richard Schwartz, Long Nguyen, and John Makhoul. 2007. Language model adaptation in machine translation from speech. *Test*, 4:117–120.

S.F. Chen and J. Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393.

Jorge Civera and Alfons Juan. 2007. Domain adaptation in statistical machine translation with mixture modelling. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

A. Clifton and A. Sarkar. 2011. Combining morpheme-based machine translation with post-processing morpheme prediction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, OR, USA*.

Adrià de de Gispert and José B. Mariño. 2008. On the impact of morphology in English to Spanish statistical MT. *Speech Communication*, 50(11-12):1034–1046.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*.

George Foster and Roland Kuhn. 2007. Mixture-Model Adaptation For SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 128–135, Stroudsburg, PA, USA. Association for Computational Linguistics.

George Foster, Roland Kuhn, and Howard Johnson. 2006. Phrasetable smoothing for statistical machine translation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 53–61, Stroudsburg, PA, USA. Association for Computational Linguistics.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 451–459, Cambridge, MA, October. Association for Computational Linguistics.

Carlos A. Henríquez Q., José B. Mariño, and Rafael E. Banchs. 2011. Deriving translation units using small additional corpora. In *Proceedings of the 15th Conference of the European Association for Machine Translation*.

Adolfo Hernández, Enric Monte, and José B. Mariño. 2010. Multiclass classification for Morphology generation in statistical machine translation. In *Proceedings of the VI Jornadas en Tecnología del Habla" and II Iberian SLTech Workshop*, pages 179–182, November. http://fala2010.uvigo.es.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876, Prague, Czech Republic, June. Association for Computational Linguistics.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit*.

Shankar Kumar and William Bryne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, Boston,MA, May 27-June 1.

Behrang Mohit, Frank Liberato, and Rebecca Hwa. 2009. Language Model Adaptation for Difficult to Translate Phrases. In *Proceedings of the 13th Annual Conference of the EAMT*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Lluís Padró, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene Castellón. 2010. Freeling 2.1: Five years of open-source language processing tools. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*, La Valletta, MALTA, May. ELRA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

John C. Platt, Nello Cristianini, and John Shawe-taylor. 2000. Large margin DAGs for multiclass classification. In *Advances in Neural Information Processing Systems*, pages 547–553. MIT Press.

Maja Popovic and Hermann Ney. 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, LREC'04*, pages 1585–1588, May.

M. Porter. 2001. Snowball: A language for stemming algorithms.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas*.

A. Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. 2008. Applying morphology generation models to machine translation. In *Proceedings of ACL-08: HLT*, pages 514–522, Columbus, Ohio, June. Association for Computational Linguistics.

Nicola Ueffing and Hermann Ney. 2003. Using pos information for statistical machine translation into morphologically rich languages. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 347–354, Stroudsburg, PA, USA. Association for Computational Linguistics.

S. Virpioja, J.J. Väyrynen, M. Creutz, and M. Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. *Machine Translation Summit XI*, 2007:491–498.

Hua Wu, Haifeng Wang, and Chengqing Zong. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 993–1000, Stroudsburg, PA, USA. Association for Computational Linguistics.