The TALP-UPC Phrase-based Translation Systems for WMT13: System Combination with Morphology Generation, Domain Adaptation and Corpus Filtering

Lluís Formiga*, Marta R. Costa-jussà*, José B. Mariño*
José A. R. Fonollosa*, Alberto Barrón-Cedeño*†, Lluís Màrquez*

*TALP Research Centre
Universitat Politècnica de Catalunya
Barcelona, Spain

†Facultad de Informática
Universidad Politécnica de Madrid
Madrid, Spain

{lluis.formiga,marta.ruiz,jose.marino,jose.fonollosa}@upc.edu
{albarron, lluism}@lsi.upc.edu

Abstract

This paper describes the TALP participation in the WMT13 evaluation campaign. Our participation is based on the combination of several statistical machine translation systems: based on standard phrase-based Moses systems. Variations include techniques such as morphology generation, training sentence filtering, and domain adaptation through unit derivation. The results show a coherent improvement on TER, METEOR, NIST, and BLEU scores when compared to our baseline system.

1 Introduction

The TALP-UPC center (Center for Language and Speech Technologies and Applications at Universitat Politècnica de Catalunya) focused on the English to Spanish translation of the WMT13 shared task.

Our primary (contrastive) run is an internal system selection comprised of different training approaches (without CommonCrawl, unless stated): (a) Moses Baseline (Koehn et al., 2007b), (b) Moses Baseline + Morphology Generation (Formiga et al., 2012b), (c) Moses Baseline + News Adaptation (Henríquez Q. et al., 2011), (d) Moses Baseline + News Adaptation + Morphology Generation , and (e) Moses Baseline + News Adaptation + Filtered CommonCrawl Adaptation (Barrón-Cedeño et al., 2013). Our secondary run includes is the full training strategy marked as *(e)* in the previous description.

The main differences with respect to our last year's participation (Formiga et al., 2012a) are: *i*) the inclusion of the CommonCrawl corpus, using a sentence filtering technique and the system combination itself, and *ii*) a system selection scheme to select the best translation among the different configurations.

The paper is organized as follows. Section 2 presents the phrase-based system and the main pipeline of our baseline system. Section 3 describes the our approaches to improve the baseline system on the English-to-Spanish task (special attention is given to the approaches that differ from last year). Section 4 presents the system combination approach once the best candidate phrase of the different subsystems are selected. Section 5 discusses the obtained results considering both internal and official test sets. Section 6 includes conclusions and further work.

2 Baseline system: Phrase-Based SMT

Our contribution is a follow up of our last year participation (Formiga et al., 2012a), based on a factored Moses from English to Spanish words plus their Part-of-Speech (POS). Factored corpora augments words with additional information, such as POS tags or lemmas. In that case, factors other than surface (e.g. POS) are usually less sparse, allowing the construction of factor-specific language models with higher-order n-grams. Such language models can help to obtain syntactically more correct outputs.

We used the standard models available in Moses as feature functions: relative frequencies, lexical weights, word and phrase penalties, *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).

As aforementioned, we developed five factored Moses-based independent systems with different

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 134–140,
Sofia, Bulgaria, August 8-9, 2013 ©2013 Association for Computational Linguistics

approaches. We explain them in Section 3. As a final decision, we applied a system selection scheme (Formiga et al., 2013; Specia et al., 2010) to consider the best candidate for each sentence, according to human trained quality estimation (QE) models. We set monotone reordering of the punctuation signs for the decoding using the Moses wall feature.

We tuned the systems using the Moses MERT (Och, 2003) implementation. Our focus was on minimizing the BLEU score (Papineni et al., 2002) of the development set. Still, for exploratory purposes, we tuned configuration *(c)* using PRO (Hopkins and May, 2011) to set the initial weights at every iteration of the MERT algorithm. However, it showed no significant differences compared to the original MERT implementation.

We trained the baseline system using all the available parallel corpora, except for common-crawl. That is, European Parliament (EPPS) (Koehn, 2005), News Commentary, and United Nations. Regarding the monolingual data, there were more News corpora organized by years for Spanish. The data is available at the Translation Task's website[1]. We used all the News corpora to busld the language model (LM). Firstly, a LM was built for every corpus independently. Afterwards, they were combined to produce de final LM.

For internal testing we used the News 2011 and News 2012 data and concatenated the remaining three years of News data as a single parallel corpus for development.

We processed the corpora as in our participation to WMT12 (Formiga et al., 2012a). Tokenization and POS-tagging in both Spanish and English was obtained with FreeLing (Padró et al., 2010). Stemming was carried out with Snowball (Porter, 2001). Words were conditionally case folded based on their POS: proper nouns and adjectives were separated from other categories to determine whether a string should be fully folded (no special property), partially folded (noun or adjective) or not folded at all in (acronym).

Bilingual corpora was filtered with the *clean-corpus-n* script of Moses (Koehn et al., 2007a), removing those pairs in which a sentence was longer than 70. For the CommonCrawl corpus we used a more complex filtering step (cf. Section 3.3).

[1] http://www.statmt.org/wmt13/translation-task.html

Postprocessing included two special scripts to recover contractions and clitics. Detruecasing was done forcing the capitals after the punctuation signs. Furthermore we used an additional script in order to check the casing of output names with respect to the source. We reused our language models and alignments (with stems) from WMT12.

3 Improvement strategies

We tried three different strategies to improve the baseline system. Section 3.1 shows a strategy based on morphology simplification plus generation. Its aim is dealing with the problems raised by morphology-rich languages, such as Spanish. Section 3.2 presents a domain–adaptation strategy that consists of deriving new units. Section 3.3 presents an advanced strategy to filter the good bi-sentences from the CommonCrawl corpus, which might be useful to perform the domain adaptation.

3.1 Morphology generation

Following the success of our WMT12 participation (Formiga et al., 2012a), our first improvement is based on the morphology generalization and generation approach (Formiga et al., 2012b). We focus our strategy on simplifying verb forms only.

The approach first translates into Spanish simplified forms (de Gispert and Mariño, 2008). The final inflected forms are predicted through a morphology generation step, based on the shallow and deep-projected linguistic information available from both source and target language sentences.

Lexical sparseness is a crucial aspect to deal with for an open-domain robust SMT when translating to morphology-rich languages (e.g. Spanish) . We knew beforehand (Formiga et al., 2012b) that morphology generalization is a good method to deal with generic translations and it provides stability to translations of the training domain.

Our morphology prediction (generation) systems are trained with the WMT13 corpora (Europarl, News, and UN) together with noisy data (OpenSubtitles). This combination helps to obtain better translations without compromising the quality of the translation models. These kind of morphology generation systems are trained with a relatively short amount of parallel data compared to standard SMT training corpora.

Our main enhancement to this strategy is the

addition of source-projected deep features to the target sentence in order to perform the morphology prediction. These features are Dependency Features and Semantic Role Labelling, obtained from the source sentence through Lund Dependency Parser[2]. These features are then projected to the target sentence as explained in (Formiga et al., 2012b).

Projected deep features are important to predict the correct verb morphology from clean and fluent text. However, the projection of deep features is sentence-fluency sensitive, making it unreliable when the baseline MT output is poor. In other words, the morphology generation strategy becomes more relevant with high-quality MT decoders, as their output is more fluent, making the shallow and deep features more reliable classifier guides.

3.2 Domain Adaptation through pivot derived units

Usually the WMT Translation Task focuses on adapting a system to a news domain, offering an in-domain parallel corpus to work with. However this corpus is relatively small compared to the other corpora. In our previous participation we demonstrated the need of performing a more aggressive domain adaptation strategy. Our strategy was based on using in-domain parallel data to adapt the translation model, but focusing on the decoding errors that the out-of-domain baseline system makes when translating the in-domain corpus.

The idea is to identify the system mistakes and use the in-domain data to learn how to correct them. To that effect, we interpolate the translation models (phrase and lexical reordering tables) with a new adapted translation model with derived units. We obtained the units identifying the mismatching parts between the non-adapted translation and the actual reference (Henríquez Q. et al., 2011). This derivation approach uses the original translation as a pivot to find a word-to-word alignment between the source side and the target correction (word-to-word alignment provided by Moses during decoding).

The word-to-word monolingual alignment between output translation target correction was obtained combining different probabilities such as *i*)lexical identity, *ii*) TER-based alignment links, *iii*) lexical model probabilities, *iv*) char-based Levenshtein distance between tokens and *v*) filtering out those alignments from NULL to a stop word ($p = -\infty$).

We empirically set the linear interpolation weight as $w = 0.60$ for the baseline translation models and $w = 0.40$ for the derived units translations models. We applied the pivot derived units strategy to the News domain and to the filtered Commoncrawl corpus (cf. Section 5). The procedure to filter out the Commoncrawl corpus is explained next.

Corpus		Sent.	Words	Vocab.	avg.len.
Original	EN	1.48M	29.44M	465.1k	19.90
	ES		31.6M	459.9k	21.45
Filtered	EN	0.78M	15.3M	278.0k	19.72
	ES		16.6M	306.8k	21.37

Table 1: Commoncrawl corpora statistics for WMT13 before and after filtering.

3.3 CommonCrawl Filtering

We used the CommonCrawl corpus, provided for the first time by the organization, as an important source of information for performing aggressive domain adaptation. To decrease the impact of the noise in the corpus, we performed an automatic pre-selection of the supposedly more correct (hence useful) sentence pairs: we applied the automatic quality estimation filters developed in the context of the FAUST project[3]. The filters' purpose is to identify cases in which the post-editions provided by casual users really improve over automatic translations.

The adaptation to the current framework is as follows. Example selection is modelled as a binary classification problem. We consider triples $(src, ref, trans)$, where *src* and *ref* stand for the source-reference sentences in the CommonCrawl corpus and *trans* is an automatic translation of the source, generated by our baseline SMT system. A triple is assigned a positive label iff *ref* is a better translation from *src* than *trans*. That is, if the translation example provided by CommonCrawl is better than the output of our baseline SMT system.

We used four feature sets to characterize the three sentences and their relationships: *surface*, *back-translation*, *noise-based* and *similarity-based*. These features try to capture (*a*) the similarity between the different texts on the basis of

[2]http://nlp.cs.lth.se/software/

[3]http://www.faust-fp7.eu

diverse measures, (*b*) the length of the different sentences (including ratios), and (*c*) the likelihood of a source or target text to include noisy text.[4] Most of them are simple, fast-calculation and language-independent features. However, back-translation features require that *trans* and *ref* are back-translated into the source language. We did it by using the TALP es-en system from WMT12.

Considering these features, we trained linear Support Vector Machines using SVM^{light} (Joachims, 1999). Our training collection was the FFF^{+} corpus, with +500 hundred manually annotated instances (Barrón-Cedeño et al., 2013). No adaptation to CommonCrawl was performed. To give an idea, classification accuracy over the test partition of the FFF^{+} corpus was only moderately good ($\sim$70%). However, ranking by classification score a fresh set of over 6,000 new examples, and selecting the top ranked 50% examples to enrich a state-of-the-art SMT system, allowed us to significantly improve translation quality (Barrón-Cedeño et al., 2013).

For WMT13, we applied these classifiers to rank the CommonCrawl translation pairs and then selected the top 53% instances to be processed by the domain adaptation strategy. Table 1 displays the corpus statistics before and after filtering.

4 System Combination

We approached system combination as a *system selection* task. More concretely, we applied Quality Estimation (QE) models (Specia et al., 2010; Formiga et al., 2013) to select the highest quality translation at sentence level among the translation candidates obtained by our different strategies. The QE models are trained with human supervision, making use of no system-dependent features.

In a previous study (Formiga et al., 2013), we showed the plausibility of building reliable system-independent QE models from human annotations. This type of task should be addressed with a pairwise ranking strategy, as it yields better results than an absolute quality estimation approach (i.e., regression) for system selection. We also found that training the quality estimation models from human assessments, instead of automatic reference scores, helped to obtain better models for system selection for both *i*) mimicking the behavior of automatic metrics and *ii*) learning the human behavior when ranking different translation candidates.

For training the QE models we used the data from the WMT13 shared task on quality estimation (*System Selection Quality Estimation at Sentence Level* task[5]), which contains the test sets from other WMT campaigns with human assessments. We used five groups of features, namely: *i*) *QuestQE*: 17 QE features provided by the Quest toolkit[6]; *ii*) *AsiyaQE*: 26 QE features provided by the Asiya toolkit for MT evaluation (Giménez and Màrquez, 2010a); *iii*) LM (and LM-PoS) perplexities trained with monolingual data; *iv*) *PR*: Classical lexical-based measures -BLEU (Papineni et al., 2002), NIST (Doddington, 2002), and METEOR (Denkowski and Lavie, 2011)- computed with a pseudo-reference approach, that is, using the other system candidates as references (Soricut and Echihabi, 2010); and *v*) *PROTHER*: Reference based metrics provided by Asiya, including GTM, ROUGE, PER, TER (Snover et al., 2008), and syntax-based evaluation measures also with a pseudo-reference approach.

We trained a Support Vector Machine ranker by means of pairwise comparison using the SVM^{light} toolkit (Joachims, 1999), but with the "`-z p`" parameter, which can provide system rankings for all the members of different groups. The learner algorithm was run according to the following parameters: linear kernel, expanding the working set by 9 variables at each iteration, for a maximum of 50,000 iterations and with a cache size of 100 for kernel evaluations. The trade-off parameter was empirically set to 0.001.

Table 2 shows the contribution of different feature groups when training the QE models. For evaluating performance, we used the Asiya normalized linear combination metric ULC (Giménez and Màrquez, 2010b), which combines BLEU, NIST, and METEOR (with exact, paraphrases and synonym variants). Within this scenario, it can be observed that the quality estimation features (QuestQE and AsiyaQE) did not obtain good results, perhaps because of the high similarity between the test candidates (Moses with different configurations) in contrast to the strong difference between the candidates in training (Moses,

[4] We refer the interested reader to (Barrón-Cedeño et al., 2013) for a detailed description of features, process, and evaluation.

[5] http://www.quest.dcs.shef.ac.uk/wmt13_qe.html

[6] http://www.quest.dcs.shef.ac.uk

Features	Asiya ULC			
	WMT'11	WMT'12	AVG	WMT'13
QuestQE	60.46	60.64	60.55	60.06
AsiyaQE	61.04	60.89	60.97	60.29
QuestQE+AsiyaQE	60.86	61.07	60.96	60.42
LM	60.84	60.63	60.74	60.37
QuestQE+AsiyaQE+LM	60.80	60.55	60.67	60.21
QuestQE+AsiyaQE+PR	60.97	61.12	61.05	60.54
QuestQE+AsiyaQE+PR+PROTHER	61.05	61.19	61.12	60.69
PR	*61.24*	61.08	61.16	***61.04***
PR+PROTHER	61.19	61.16	61.18	60.98
PR+PROTHER+LM	61.11	*61.29*	***61.20***	61.03
QuestQE+AsiyaQE+PR+PROTHER+LM	60.70	60.88	60.79	60.14

Table 2: System selection scores (ULC) obtained using QE models trained with different groups of features. Results displayed for WMT11, WMT12 internal tests, their average, and the WMT13 test

EN→ES		BLEU	TER
wmt13	Primary	29.5	0.586
wmt13	Secondary	29.4	0.586

Table 4: Official automatic scores for the WMT13 English↔Spanish translations.

RBMT, Jane, etc.). On the contrary, the pseudo-reference-based features play a crucial role in the proper performance of the QE model, confirming the hypothesis that PR features need a clear dominant system to be used as reference. The PR-based configurations (with and without LM) had no big differences between them. We choose the best AVG result for the final system combination: *PR+PROTHER+LM*, which it is consistent with the actual WMT13 evaluated afterwards.

5 Results

Evaluations were performed considering different quality measures: BLEU, NIST, TER, and METEOR in addition to an informal manual analysis. This manifold of metrics evaluates distinct aspects of the translation. We evaluated both over the WMT11 and WMT12 test sets as internal indicators of our systems. We also give our performance on the WMT13 test dataset.

Table 3 presents the obtained results for the different strategies: (a) Moses Baseline (w/o commoncrawl) (b) Moses Baseline+Morphology Generation (w/o commoncrawl) (c) Moses Baseline+News Adaptation through pivot based alignment (w/o commoncrawl) (d) Moses Baseline + News Adaptation (b) + Morphology Generation (c) (e) Moses Baseline + News Adaptation (b) + Filtered CommonCrawl Adaptation.

The official results are in Table 4. Our primary (contrastive) run is the system combination strategy whereas our secondary run is the full training strategy marked as (e) on the system combination. Our primary system was ranked in the second cluster out of ten constrained systems in the official manual evaluation.

Independent analyzes of the improvement strategies show that the highest improvement comes from the CommonCrawl Filtering + Adaptation strategy (system e). The second best strategy is the combination of the morphology prediction system plus the news adaptation system. However, for the WMT12 test the News Adaptation strategy contributes to main improvement whereas for the WMT13 this major improvement is achieved with the morphology strategy. Analyzing the distance betweem each test set with respect to the News and CommonCrawl domain to further understand the behavior of each strategy seems an interesting future work. Specifically, for further contrasting the difference in the morphology approach, it would be nice to analyze the variation in the verb inflection forms. Hypothetically, the person or the number of the verb forms used may have a higher tendency to be different in the WMT13 test set, implying that our morphology approach is further exploited.

Regarding the system selection step (internal WMT12 test), the only automatic metric that has an improvement is TER. However, TER is one of

EN→ES		BLEU	NIST	TER	METEOR
wmt12	Baseline	32.97	8.27	49.27	49.91
wmt12	+ Morphology Generation	33.03	8.29	49.02	50.01
wmt12	+ News Adaptation	33.22	8.31	49.00	50.16
wmt12	+ News Adaptation + Morphology Generation	33.29	8.32	48.83	50.29
wmt12	+ News Adaptation + Filtered CommonCrawl Adaptation	**33.61**	**8.35**	48.82	**50.52**
wmt12	System Combination	33.43	8.34	**48.78**	50.44
wmt13	Baseline	29.02	7.72	51.92	46.96
wmt13	Morphology Generation	29.35	7.73	52.04	47.04
wmt13	News Adaptation	29.19	7.74	51.91	47.07
wmt13	News Adaptation + Morphology Generation	29.40	7.74	51.96	47.12
wmt13	News Adaptation + Filtered CommonCrawl Adaptation	29.47	7.77	51.82	47.22
wmt13	System Combination	**29.54**	**7.77**	**51.76**	**47.34**

Table 3: Automatic scores for English→Spanish translations.

the most reliable metrics according to human evaluation. Regarding the actual WMT13 test, the system selection step is able to overcome all the automatic metrics.

6 Conclusions and further work

This paper described the TALP-UPC participation for the English-to-Spanish WMT13 translation task. We applied the same systems as in last year, but enhanced with new techniques: sentence filtering and system combination.

Results showed that both approaches performed better than the baseline system, being the sentence filtering technique the one that most improvement reached in terms of all the automatic quality indicators: BLEU, NIST, TER, and METEOR. The system combination was able to outperform the independent systems which used morphological knowledge and/or domain adaptation techniques.

As further work would like to focus on further advancing on the morphology-based techniques.

Acknowledgments

This work has been supported in part by Spanish Ministerio de Economía y Competitividad, contract TEC2012-38939-C03-02 as well as from the European Regional Development Fund (ERDF/FEDER) and the European Community's FP7 (2007-2013) program under the following grants: 247762 (FAUST, FP7-ICT-2009-4-247762), 29951 (the International Outgoing Fellowship Marie Curie Action – IMTraP-2011-29951) and 246016 (ERCIM "Alain Bensoussan" Fellowship).

References

Alberto Barrón-Cedeño, Lluís Màrquez, Carlos A. Henríquez Q, Lluís Formiga, Enrique Romero, and Jonathan May. 2013. Identifying Useful Human Correction Feedback from an On-line Machine Translation Service. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*. AAAI Press.

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 Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lluis Formiga, Carlos A. Henríquez Q., Adolfo Hernández, José B. Mariño, Enric Monte, and José A. R. Fonollosa. 2012a. The TALP-UPC phrase-based translation systems for WMT12: Morphology simplification and domain adaptation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 275–282, Montréal, Canada, June. Association for Computational Linguistics.

Lluís Formiga, Adolfo Hernández, José B. Mariñ, and Enrique Monte. 2012b. Improving english to spanish out-of-domain translations by morphology generalization and generation. In *Proceedings of*

the AMTA Monolingual Machine Translation-2012 Workshop.

Lluís Formiga, Lluís Màrquez, and Jaume Pujantell. 2013. Real-life translation quality estimation for mt system selection. In *Proceedings of 14th Machine Translation Summit (MT Summit)*, Nice, France, September. EAMT.

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.

Jesús Giménez and Lluís Màrquez. 2010a. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86.

Jesús Giménez and Lluís Màrquez. 2010b. Linguistic measures for automatic machine translation evaluation. *Machine Translation*, 24(3-4):209–240, December.

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*.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Thorsten Joachims, 1999. *Advances in Kernel Methods – Support Vector Learning*, chapter Making large-Scale SVM Learning Practical. MIT Press.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. 2007a. 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, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007b. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

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

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)*.

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

Matthew Snover, Bonnie Dorr, and Richard Schwartz. 2008. Language and Translation Model Adaptation using Comparable Corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*.

Radu Soricut and Abdessamad Echihabi. 2010. Trustrank: Inducing trust in automatic translations via ranking. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 612–621, Uppsala, Sweden, July. Association for Computational Linguistics.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine Translation Evaluation Versus Quality Estimation. *Machine Translation*, 24:39–50, March.