Abu-MaTran at WMT 2014 Translation Task: Two-step Data Selection and RBMT-Style Synthetic Rules

Raphael Rubino[⋆], Antonio Toral[†], Victor M. Sánchez-Cartagena[⋆‡], Jorge Ferrández-Tordera[⋆], Sergio Ortiz-Rojas[⋆], Gema Ramírez-Sánchez[⋆], Felipe Sánchez-Martínez[‡], Andy Way[†]

[⋆] Prompsit Language Engineering, S.L., Elche, Spain
{rrubino,vmsanchez,jferrandez,sortiz,gramirez}@prompsit.com
[†] NCLT, School of Computing, Dublin City University, Ireland
{atoral,away}@computing.dcu.ie
[‡] Dep. Llenguatges i Sistemes Informàtics, Universitat d'Alacant, Spain
fsanchez@dlsi.ua.es

Abstract

This paper presents the machine translation systems submitted by the Abu-MaTran project to the WMT 2014 translation task. The language pair concerned is English–French with a focus on French as the target language. The French to English translation direction is also considered, based on the word alignment computed in the other direction. Large language and translation models are built using all the datasets provided by the shared task organisers, as well as the monolingual data from LDC. To build the translation models, we apply a two-step data selection method based on bilingual cross-entropy difference and vocabulary saturation, considering each parallel corpus individually. Synthetic translation rules are extracted from the development sets and used to train another translation model. We then interpolate the translation models, minimising the perplexity on the development sets, to obtain our final SMT system. Our submission for the English to French translation task was ranked second amongst nine teams and a total of twenty submissions.

1 Introduction

This paper presents the systems submitted by the Abu-MaTran project (runs named *DCU-Prompsit-UA*) to the WMT 2014 translation task for the English–French language pair. Phrase-based statistical machine translation (SMT) systems were submitted, considering the two translation directions, with the focus on the English to French direction. Language models (LMs) and translation models (TMs) are trained using all the data provided by the shared task organisers, as well as the *Gigaword* monolingual corpora distributed by LDC.

To train the LMs, monolingual corpora and the target side of the parallel corpora are first used individually to train models. Then the individual models are interpolated according to perplexity minimisation on the development sets.

To train the TMs, first a baseline is built using the *News Commentary* parallel corpus. Second, each remaining parallel corpus is processed individually using bilingual cross-entropy difference (Axelrod et al., 2011) in order to separate *pseudo* in-domain and out-of-domain sentence pairs, and filtering the *pseudo* out-of-domain instances with the vocabulary saturation approach (Lewis and Eetemadi, 2013). Third, synthetic translation rules are automatically extracted from the development set and used to train another translation model following a novel approach (Sánchez-Cartagena et al., 2014). Finally, we interpolate the four translation models (baseline, in-domain, filtered out-of-domain and rules) by minimising the perplexity obtained on the development sets and investigate the best tuning and decoding parameters.

The reminder of this paper is organised as follows: the datasets and tools used in our experiments are described in Section 2. Then, details about the LMs and TMs are given in Section 3 and Section 4 respectively. Finally, we evaluate the performance of the final SMT system according to different tuning and decoding parameters in Section 5 before presenting conclusions in Section 6.

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 171–177,
Baltimore, Maryland USA, June 26–27, 2014.

2 Datasets and Tools

We use all the monolingual and parallel datasets in English and French provided by the shared task organisers, as well as the LDC *Gigaword* for the same languages[1]. For each language, a true-case model is trained using all the data, using the *train-truecaser.perl* script included in the MOSES toolkit (Koehn et al., 2007).

Punctuation marks of all the monolingual and parallel corpora are then normalised using the script *normalize-punctuation.perl* provided by the organisers, before being tokenised and true-cased using the scripts distributed with the MOSES toolkit. The same pre-processing steps are applied to the development and test sets. As development sets, we used all the test sets from previous years of WMT, from 2008 to 2013 (*newstest2008-2013*).

Finally, the training parallel corpora are cleaned using the script *clean-corpus-n.perl*, keeping the sentences longer than 1 word, shorter than 80 words, and with a length ratio between sentence pairs lower than 4.[2] The statistics about the corpora used in our experiments after pre-processing are presented in Table 1.

For training LMs we use KENLM (Heafield et al., 2013) and the SRILM tool-kit (Stolcke et al., 2011). For training TMs, we use MOSES (Koehn et al., 2007) version 2.1 with MGIZA++ (Och and Ney, 2003; Gao and Vogel, 2008). These tools are used with default parameters for our experiments except when explicitly said.

The decoder used to generate translations is MOSES using features weights optimised with MERT (Och, 2003). As our approach relies on training individual TMs, one for each parallel corpus, our final TM is obtained by linearly interpolating the individual ones. The interpolation of TMs is performed using the script *tmcombine.py*, minimising the cross-entropy between the TM and the concatenated development sets from 2008 to 2012 (noted *newstest2008-2012*), as described in Sennrich (2012). Finally, we make use of the findings from WMT 2013 brought by the winning team (Durrani et al., 2013) and decide to use the Operation Sequence Model (OSM), based on minimal translation units and Markov chains over sequences of operations, implemented in MOSES and introduced by Durrani et al. (2011).

[1]LDC2011T07 English *Gigaword Fifth Edition*, LDC2011T10 French *Gigaword Third Edition*

[2]This ratio was empirically chosen based on words fertility between English and French.

Corpus	Sentences (k)	Words (M)
Monolingual Data – English		
Europarl v7	2,218.2	59.9
News Commentary v8	304.2	7.4
News Shuffled 2007	3,782.5	90.2
News Shuffled 2008	12,954.5	308.1
News Shuffled 2009	14,680.0	347.0
News Shuffled 2010	6,797.2	157.8
News Shuffled 2011	15,437.7	358.1
News Shuffled 2012	14,869.7	345.5
News Shuffled 2013	21,688.4	495.2
LDC afp	7,184.9	869.5
LDC apw	8,829.4	1,426.7
LDC cna	618.4	45.7
LDC ltw	986.9	321.1
LDC nyt	5,327.7	1,723.9
LDC wpb	108.8	20.8
LDC xin	5,121.9	423.7
Monolingual Data – French		
Europarl v7	2,190.6	63.5
News Commentary v8	227.0	6.5
News Shuffled 2007	119.0	2.7
News Shuffled 2008	4,718.8	110.3
News Shuffled 2009	4,366.7	105.3
News Shuffled 2010	1,846.5	44.8
News Shuffled 2011	6,030.1	146.1
News Shuffled 2012	4,114.4	100.8
News Shuffled 2013	9,256.3	220.2
LDC afp	6,793.5	784.5
LDC apw	2,525.1	271.3
Parallel Data		
10^9 Corpus	21,327.1	549.0 (EN) 642.5 (FR)
Common Crawl	3,168.5	76.0 (EN) 82.7 (FR)
Europarl v7	1,965.5	52.5 (EN) 56.7 (FR)
News Commentary v9	181.3	4.5 (EN) 5.3 (FR)
UN	12,354.7	313.4 (EN) 356.5 (FR)

Table 1: Data statistics after pre-processing of the monolingual and parallel corpora used in our experiments.

3 Language Models

The LMs are trained in the same way for both languages. First, each monolingual and parallel corpus is considered individually (except the parallel version of Europarl and News Commentary) and used to train a 5-gram LM with the modified Kneser-Ney smoothing method. We then interpolate the individual LMs using the script *compute-best-mix* available with the SRILM tool-kit (Stolcke et al., 2011), based on their perplexity scores on the concatenation of the development sets from 2008 to 2012 (the 2013 version is held-out for the tuning of the TMs).

The final LM for French contains all the word sequences from 1 to 5-grams contained in the training corpora without any pruning. However, with the computing resources at our disposal, the English LMs could not be interpolated without pruning non-frequent n-grams. Thus, n-grams with $n \in [3;5]$ with a frequency lower than 2 were removed. Details about the final LMs are given in Table 2.

	1-gram	2-gram	3-gram	4-gram	5-gram
English	13.4	198.6	381.2	776.3	1,068.7
French	6.0	75.5	353.2	850.8	1,354.0

Table 2: Statistics, in millions of n-grams, of the interpolated LMs.

4 Translation Models

In this Section, we describe the TMs trained for the shared task. First, we present the two-step data selection process which aims to (i) separate in and out-of-domain parallel sentences and (ii) reduce the total amount of out-of-domain data. Second, a novel approach for the automatic extraction of translation rules and their use to enrich the phrase table is detailed.

4.1 Parallel Data Filtering and Vocabulary Saturation

Amongst the parallel corpora provided by the shared task organisers, only *News Commentary* can be considered as in-domain regarding the development and test sets. We use this training corpus to build our baseline SMT system. The other parallel corpora are individually filtered using bilingual cross-entropy difference (Moore and Lewis, 2010; Axelrod et al., 2011). This data filtering method relies on four LMs, two in the source and two in the target language, which aim to model particular features of in and out-of-domain sentences.

We build the in-domain LMs using the source and target sides of the *News Commentary* parallel corpus. Out-of-domain LMs are trained on a vocabulary-constrained subset of each remaining parallel corpus individually using the SRILM toolkit, which leads to eight models (four in the source language and four in the target language).[3]

[3]The subsets contain the same number of sentences and the same vocabulary as *News Commentary*.

Then, for each out-of-domain parallel corpus, we compute the bilingual cross-entropy difference of each sentence pair as:

$$[H_{\text{in}}(S_{\text{src}}) - H_{\text{out}}(S_{\text{src}})] + [H_{\text{in}}(S_{\text{trg}}) - H_{\text{out}}(S_{\text{trg}})] \quad (1)$$

where S_{src} and S_{trg} are the source and the target sides of a sentence pair, H_{in} and H_{out} are the cross-entropies of the in and out-of-domain LMs given a sentence pair. The sentence pairs are then ranked and the lowest-scoring ones are taken to train the *pseudo* in-domain TMs. However, the cross-entropy difference threshold required to split a corpus in two parts (*pseudo* in and out-of-domain) is usually set empirically by testing several subset sizes of the top-ranked sentence pairs. This method is costly in our setup as it would lead to training and evaluating multiple SMT systems for each of the *pseudo* in-domain parallel corpora.

In order to save time and computing power, we consider only *pseudo* in-domain sentence pairs those with a bilingual cross-entropy difference below 0, i.e. those deemed more similar to the in-domain LMs than to the out-of-domain LMs ($H_{\text{in}} < H_{\text{out}}$). A sample of the distribution of scores for the out-of-domain corpora is shown in Figure 1. The resulting *pseudo* in-domain corpora are used to train individual TMs, as detailed in Table 3.

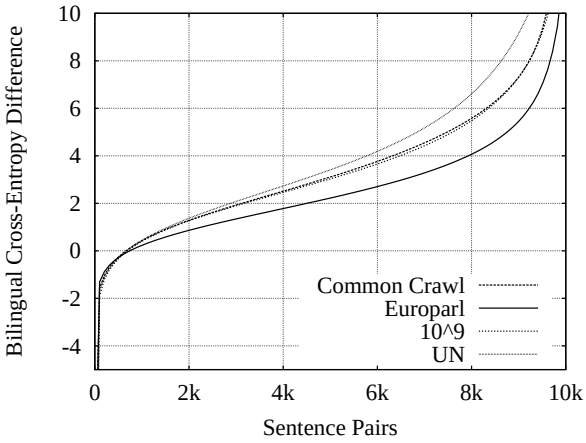

Figure 1: Sample of ranked sentence-pairs (10k) of each of the out-of-domain parallel corpora with bilingual cross-entropy difference

The results obtained using the *pseudo* in-domain data show BLEU (Papineni et al., 2002) scores superior or equal to the baseline score. Only the *Europarl* subset is slightly lower than the baseline, while the subset taken from the 10^9 corpus reaches the highest BLEU compared to the other systems (30.29). This is mainly due to the

size of this subset which is ten times larger than the one taken from *Europarl*. The last row of Table 3 shows the BLEU score obtained after interpolating the four *pseudo* in-domain translation models. This system outperforms the best *pseudo* in-domain one by 0.5 absolute points.

Corpus	Sentences (k)	BLEU_{dev}
Baseline	181.3	27.76
Common Crawl	208.3	27.73
Europarl	142.0	27.63
10^9 Corpus	1,442.4	30.29
UN	642.4	28.91
Interpolation	-	30.78

Table 3: Number of sentence pairs and BLEU scores reported by MERT on English–French *newstest2013* for the *pseudo* in-domain corpora obtained by filtering the out-of-domain corpora with bilingual cross-entropy difference. The interpolation of *pseudo* in-domain models is evaluated in the last row.

After evaluating the *pseudo* in-domain parallel data, the remaining sentence pairs for each corpora are considered out-of-domain according to our filtering approach. However, they may still contain useful information, thus we make use of these corpora by building individual TMs for each corpus (in a similar way we built the *pseudo* in-domain models). The total amount of remaining data (more than 33 million sentence pairs) makes the training process costly in terms of time and computing power. In order to reduce these costs, sentence pairs with a bilingual cross-entropy difference higher than 10 were filtered out, as we noticed that most of the sentences above this threshold contain noise (non-alphanumeric characters, foreign languages, etc.).

We also limit the size of the remaining data by applying the vocabulary saturation method (Lewis and Eetemadi, 2013). For the out-of-domain subset of each corpus, we traverse the sentence pairs in the order they are ranked by perplexity difference and filter out those sentence pairs for which we have seen already each 1-gram at least 10 times. Each out-of-domain subset from each parallel corpus is then used to train a TM before interpolating them to create the *pseudo* out-of-domain TM. The results reported by MERT obtained on the *newstest2013* development set are detailed in Table 4.

Mainly due to the sizes of the *pseudo* out-of-domain subsets, the reported BLEU scores are higher than the baseline for the four individual SMT systems and the interpolated one. This latter system outperforms the baseline by 3.61 absolute points. Compared to the results obtained with the *pseudo* in-domain data, we observe a slight improvement of the BLEU scores using the *pseudo* out-of-domain data. However, despite the comparatively larger sizes of the latter datasets, the BLEU scores reached are not that higher. For instance with the 10^9 corpus, the *pseudo* in and out-of-domain subsets contain 1.4 and 5.1 million sentence pairs respectively, and the two systems reach 30.3 and 30.5 BLEU. These scores indicate that the *pseudo* in-domain SMT systems are more efficient on the English–French *newstest2013* development set.

Corpus	Sentences (k)	BLEU_{dev}
Baseline	181.3	27.76
Common Crawl	1,598.7	29.84
Europarl	461.9	28.87
10^9 Corpus	5,153.0	30.50
UN	1,707.3	29.03
Interpolation	-	31.37

Table 4: Number of sentence pairs and BLEU scores reported by MERT on English–French *newstest2013* for the *pseudo* out-of-domain corpora obtained by filtering the out-of-domain corpora with bilingual cross-entropy difference, keeping sentence pairs below an entropy score of 10 and applying vocabulary saturation. The interpolation of *pseudo* out-of-domain models is evaluated in the last row.

4.2 Extraction of Translation Rules

A synthetic phrase-table based on shallow-transfer MT rules and dictionaries is built as follows. First, a set of shallow-transfer rules is inferred from the concatenation of the *newstest2008-2012* development corpora exactly in the same way as in the UA-Prompsit submission to this translation shared task (Sánchez-Cartagena et al., 2014). In summary, rules are obtained from a set of bilingual phrases extracted from the parallel corpus after its morphological analysis and part-of-speech disambiguation with the tools in the Apertium rule-based MT platform (Forcada et al., 2011).

The extraction algorithm commonly used in phrase-based SMT is followed with some added heuristics which ensure that the bilingual phrases

extracted are compatible with the bilingual dictionary. Then, many different rules are generated from each bilingual phrase; each of them encodes a different degree of generalisation over the particular example it has been extracted from. Finally, the minimum set of rules which correctly reproduces all the bilingual phrases is found based on integer linear programming search (Garfinkel and Nemhauser, 1972).

Once the rules have been inferred, the phrase table is built from them and the original rule-based MT dictionaries, following the method by Sánchez-Cartagena et al. (2011), which was one of winning systems[4] (together with two online SMT systems) in the pairwise manual evaluation of the WMT11 English–Spanish translation task (Callison-Burch et al., 2011). This phrase-table is then interpolated with the baseline TM and the results are presented in Table 5. A slight improvement over the baseline is observed, which motivates the use of synthetic rules in our final MT system. This small improvement may be related to the small coverage of the Apertium dictionaries: the English–French bilingual dictionary has a low number of entries compared to more mature language pairs in Apertium which have around 20 times more bilingual entries.

System	BLEU_{dev}
Baseline	27.76
Baseline+Rules	28.06

Table 5: BLEU scores reported by MERT on English–French *newstest2013* for the baseline SMT system standalone and with automatically extracted translation rules.

5 Tuning and Decoding

We present in this Section a short selection of our experiments, amongst 15+ different configurations, conducted on the interpolation of TMs, tuning and decoding parameters. We first interpolate the four TMs: the baseline, the *pseudo* in and out-of-domain, and the translation rules, minimising the perplexity obtained on the concatenated development sets from 2008 to 2012 (*newstest2008-2012*). We investigate the use of OSM trained on *pseudo* in-domain data only or using all the parallel data available. Finally, we make variations of the number of n-bests used by MERT.

Results obtained on the development set *newstest2013* are reported in Table 6. These scores show that adding OSM to the interpolated translation models slightly degrades BLEU. However, by increasing the number of n-bests considered by MERT to 200-bests, the SMT system with OSM outperforms the systems evaluated previously in our experiments. Adding the synthetic translation rules degrades BLEU (as indicated by the last row in the Table), thus we decide to submit two systems to the shared task: one without and one with synthetic rules. By submitting a system without synthetic rules, we also ensure that our SMT system is constrained according to the shared task guidelines.

System	BLEU_{dev}
Baseline	27.76
+ *pseudo* in + *pseudo* out	31.93
+ OSM	31.90
+ MERT 200-best	32.21
+ Rules	32.10

Table 6: BLEU scores reported by MERT on English–French *newstest2013* development set.

As MERT is not suitable when a large number of features are used (our system uses 19 fetures), we switch to the Margin Infused Relaxed Algorithm (MIRA) for our submitted systems (Watanabe et al., 2007). The development set used is *newstest2012*, as we aim to select the best decoding parameters according to the scores obtained when decoding the *newstest2013* corpus, after detruecasing and de-tokenising using the scripts distributed with MOSES. This setup allowed us to compare our results with the participants of the translation shared task last year. We pick the decoding parameters leading to the best results in terms of BLEU and decode the official test set of WMT14 *newstest2014*. The results are reported in Table 7. Results on *newstest2013* show that the decoding parameters investigation leads to an overall improvement of 0.1 BLEU absolute. The results on *newstest2014* show that adding synthetic rules did not help improving BLEU and degraded slightly TER (Snover et al., 2006) scores.

In addition to our English→French submission, we submitted a French→English translation. Our French→English MT system is built on the alignments obtained from the English→French direction. The training processes between the two sys-

[4]No other system was found statistically significantly better using the sign test at $p \leq 0.1$.

System	BLEU13A	TER
newstest2013		
Best tuning	31.02	60.77
cube-pruning (pop-limit 10000)	31.04	60.71
increased table-limit (100)	31.06	60.77
monotonic reordering	31.07	60.69
Best decoding	31.14	60.66
newstest2014		
Best decoding	34.90	54.70
Best decoding + Rules	34.90	54.80

Table 7: Case sensitive results obtained with our final English–French SMT system on *newstest2013* when experimenting with different decoding parameters. The best parameters are kept to translate the WMT14 test set (*newstest2014*) and official results are reported in the last two rows.

tems are identical, except for the synthetic rules which are not extracted for the French→English direction. Tuning and decoding parameters for this latter translation direction are the best ones obtained in our previous experiments on this shared task. The case-sensitive scores obtained for French→English on *newstest2014* are 35.0 BLEU13A and 53.1 TER, which ranks us at the fifth position for this translation direction.

6 Conclusion

We have presented the MT systems developed by the Abu-MaTran project for the WMT14 translation shared task. We focused on the French–English language pair and particularly on the English→French direction. We have used a two-step data selection process based on bilingual cross-entropy difference and vocabulary saturation, as well as a novel approach for the extraction of synthetic translation rules and their use to enrich the phrase table. For the LMs and the TMs, we rely on training individual models per corpus before interpolating them by minimising perplexity according to the development set. Finally, we made use of the findings of WMT13 by including an OSM model.

Our English→French translation system was ranked second amongst nine teams and a total of twenty submissions, while our French→English submission was ranked fifth. As future work, we plan to investigate the effect of adding to the phrase table synthetic translation rules based on larger dictionaries. We also would like to study the link between OSM and the different decoding parameters implemented in MOSES, as we observed inconsistent results in our experiments.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran).

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain Adaptation Via Pseudo In-domain Data Selection. In *Proceedings of EMNLP*, pages 355–362.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of WMT*, pages 22–64.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A Joint Sequence Translation Model with Integrated Reordering. In *Proceedings of ACL/HLT*, pages 1045–1054.

Nadir Durrani, Barry Haddow, Kenneth Heafield, and Philipp Koehn. 2013. Edinburgh's Machine Translation Systems for European Language Pairs. In *Proceedings of WMT*, pages 112–119.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: A Free/Open-source Platform for Rule-based Machine Translation. *Machine Translation*, 25(2):127–144.

Qin Gao and Stephan Vogel. 2008. Parallel Implementations of Word Alignment Tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57.

Robert S Garfinkel and George L Nemhauser. 1972. *Integer Programming*, volume 4. Wiley New York.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable Modified Kneser-Ney Language Model Estimation. In *Proceedings of ACL*, pages 690–696.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL, Interactive Poster and Demonstration Sessions*, pages 177–180.

William D. Lewis and Sauleh Eetemadi. 2013. Dramatically Reducing Training Data Size Through Vocabulary Saturation. In *Proceedings of WMT*, pages 281–291.

Robert C. Moore and William Lewis. 2010. Intelligent Selection of Language Model Training Data. In *Proceedings of ACL*, pages 220–224.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29:19–51.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of ACL*, volume 1, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*, pages 311–318.

Víctor M. Sánchez-Cartagena, Felipe Sánchez-Martínez, and Juan Antonio Pérez-Ortiz. 2011. Integrating Shallow-transfer Rules into Phrase-based Statistical Machine Translation. In *Proceedings of MT Summit XIII*, pages 562–569.

Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2014. The UA-Prompsit Hybrid Machine Translation System for the 2014 Workshop on Statistical Machine Translation. In *Proceedings of WMT*.

Rico Sennrich. 2012. Perplexity Minimization for Translation Model Domain Adaptation in Statistical Machine Translation. In *Proceedings of EACL*, pages 539–549.

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 AMTA*, pages 223–231.

Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. 2011. SRILM at Sixteen: Update and Outlook. In *Proceedings of ASRU*.

Taro Watanabe, Jun Suzuki, Hajime Tsukada, and Hideki Isozaki. 2007. Online Large-margin Training for Statistical Machine Translation. In *Proceedings of EMNLP*.