UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Improving Statistical Machine Translation Performance by Oracle-BLEU Model Re-estimation

Dakwale, P.; Monz, C.

DOI
10.18653/v1/P16-2007

Publication date
2016

Document Version
Final published version

Published in
The 54th Annual Meeting of the Association for Computational Linguistics : ACL 2016

License
CC BY

Link to publication

Citation for published version (APA):
Dakwale, P., & Monz, C. (2016). Improving Statistical Machine Translation Performance by Oracle-BLEU Model Re-estimation. In K. Erk, & N. A. Smith (Eds.), *The 54th Annual Meeting of the Association for Computational Linguistics : ACL 2016: proceedings of the conference : August 7-12, 2016, Berlin Germany* (Vol. 2, pp. 38-44). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-2007

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:08 Dec 2021

Improving Statistical Machine Translation Performance by Oracle-BLEU Model Re-estimation

Praveen Dakwale
Informatics Institute
University of Amsterdam
p.dakwale@uva.nl

Christof Monz
Informatics Institute
University of Amsterdam
c.monz@uva.nl

Abstract

We present a novel technique for training translation models for statistical machine translation by aligning source sentences to their oracle-BLEU translations. In contrast to previous approaches which are constrained to phrase training, our method also allows the re-estimation of reordering models along with the translation model. Experiments show an improvement of up to 0.8 BLEU for our approach over a competitive Arabic-English baseline trained directly on the word-aligned bitext using heuristic extraction. As an additional benefit, the phrase table size is reduced dramatically to only 3% of the original size.

1 Introduction

In phrase-based SMT, the phrase pairs in the translation model are traditionally trained by applying a heuristic extraction method (Och and Ney, 2000) which extracts phrase pairs based on consistency of word alignments from a word-aligned bilingual training data. The probabilities of the translation model are then calculated based on the relative frequencies of the extracted phrase pairs.

A notable shortcoming of this approach is that the translation model probabilities thus calculated from the training bitext can be unintuitive and unreliable (Marcu and Wong, 2002; Foster et al., 2006) as they reflect only the distribution over the phrase pairs observed in the training data.

However, from an SMT perspective it is important that the models reflect probability distributions which are preferred by the decoding process, i.e., phrase translations which are likely to be used frequently to achieve better translations should get higher scores and phrases which are less likely to be used should get low scores. In addition, the heuristic extraction algorithm generates all possible, consistent phrases including overlapping phrases. This means that translation probabilities are distributed over a very large number of phrase translation candidates most of which never lead to the best possible translation of a sentence.

In this paper, we propose a novel solution which is to re-estimate the models from the best BLEU translation of each source sentence in the bitext. An important contribution of our approach is that unlike previous approaches such as forced alignment (Wuebker et al., 2010), reordering and language models can also be re-estimated.

2 Related Work

The forced alignment technique of Wuebker et al. (2010) forms the main motivation for our work. In forced alignment, given a sentence pair (F, E), a decoder determines the best phrase segmentation and alignment which will result in a translation of F into E. The best segmentation is defined as the one which maximizes the probability of translating the source sentence into the given target sentence. At the end, the phrase table is re-estimated using the phrase pair segmentations obtained from forced decoding. Thus forced alignment is a re-estimation technique where translation probabilities are calculated based on their frequency in best-scoring hypotheses instead of the frequencies of all possible phrase pairs in the bitext. However, one limitation of forced alignment is that only the phrase translation model can be re-estimated since it is restricted to align the source sentence to the given target reference, thus fixing the choice of reordering decisions.

A similar line of work is proposed by Lambert et al. (2011) and Schwenk et al. (2011) who use a self-enhancing strategy to utilize additional mono-

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 38–44,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

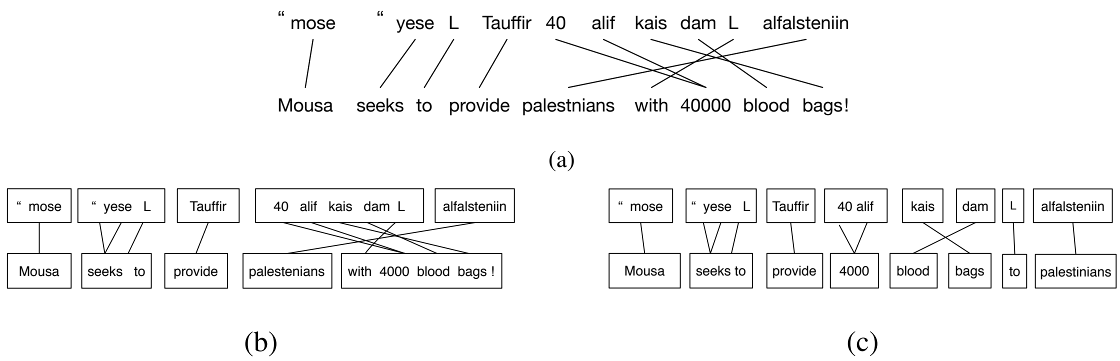

Figure 1: (a) : Word alignment from EM training for Arabic (transliterated) -English sentence pair. (b): Phrase segmentations and alignments from forced decoding. (c): Phrase segmentations and alignments from oracle BLEU re-estimation. Blocks represent phrase boundaries.

lingual source language data by aligning it to its target language translation obtained by using an SMT system to rank sentence translation probabilities. However, the main focus of their work is translation model adaptation by augmenting the bitext with additional training data and not the re-estimation of the translation models trained on the parallel data.

In this work, we propose that aligning source sentences to their oracle BLEU translations provides a more realistic estimate of the models from the decoding perspective instead of aligning them to high quality human translations as in forced decoding.

Another relevant line of research relates tuning (weight optimisation), where our work lies between forced decoding (Wuebker et al., 2010) and the bold updating approach of (Liang et al., 2006). However, our approach specifically proposes a novel method for training models using oracle BLEU translations.

3 Model Re-estimation

The idea of our approach is to re-estimate the models with n-best oracle-BLEU translations and sentence alignments resulting from decoding the source sentence. Given a source and its reference translation, the oracle-BLEU translation is defined as the translation output with highest BLEU score. Oracle BLEU translations have been previously used for different analytical purposes in SMT (Srivastava et al., 2011; Dreyer et al., 2007; Wisniewski et al., 2010).

Figure 1 shows example of word alignment obtained from EM training, segmentations and alignment obtained from forced decoding and oracle-BLEU re-estimation.

3.1 Oracle BLEU

Ideally, one would like to re-estimate translation models directly from the n-best BLEU translations. However there are two problems in calculating BLEU for individual sentence: First, as discussed in (Chiang et al., 2008), BLEU is not designed to be used for sentences in isolation where it can exhibit rather volatile behavior. Hence, following their work and (Watanabe et al., 2007), we calculate BLEU for a sentence in the context of a exponentially-weighted moving average of previous translations. We briefly discuss the computation from (Chiang et al., 2008) as follows: Given a source sentence **f**, and its reference translation **r**, for an n-best translation e^*, let c(e) be defined as the vector of target length |e|, source length |f|, reference length |r|, and the number of n-gram matches between **e** and **r**, then two pseudo document parameters $\mathbf{O}$ and $\mathbf{O}_f$ are defined as:

$$\mathbf{O} \leftarrow 0.9 \cdot (\mathbf{O} + c(e^*)), \mathbf{O}_f \leftarrow 0.9 \cdot (\mathbf{O}_f + |f|) \quad (1)$$

$\mathbf{O}$ is an exponentially-weighted moving average of the vectors from previous sentences and $\mathbf{O}_f$ is the correction of source length with respect to the previous sentences. Then the BLEU score for a sentence pairs (f,r) and translation e^* is defined as:

$$\mathbf{B}(e; f, r) = (\mathbf{O}_f + |f|) \cdot BLEU(O + c(e^*; r)) \quad (2)$$

The second problem as discussed in Chiang et al. (2008) is that due to noise in the training data, a high-BLEU translation may contain certain rules which are unlikely to be used by the model. Hence following them, we use a weighted combination of BLEU and model score to select the n-best list:

$$e^* = argmax_e(B(e) - \mu \cdot (B(e) - h(e).w)) \quad (3)$$

where B(e) and h(e) are the BLEU and model scores of the candidate translation and w is the optimised weights for the models, μ controls the preference between BLEU and model scores to determine oracle translations. We set μ=0.5 to balance between BLEU scores almost as high as the max-BLEU translations, while staying close to translations preferred by the model. We also conducted a set of experiments with μ=0 (pure or absolute BLEU) in order to verify the necessity for the optimal combination. The lower scores for this setting as compared to the baseline verified that using only the best BLEU translation indeed degrades the performance of the re-estimated models. This finding for the optimal value of μ has also been established in (Chiang et al., 2008) through a series of experiments.

3.2 Training

For obtaining the oracle-BLEU translations, we first train the translation models from the bitext using the standard pipeline of word alignment and heuristic extraction. Along with the phrase translation and language models, we also train a bilingual language model (BiLM) (Niehues et al., 2011; Garmash and Monz, 2014), as well as lexicalized (Tillman, 2004) and hierarchical reordering models (Galley and Manning, 2008). We use a BiLM specifically as an instance of a reordering model in order to determine the effect of re-estimating re-ordering decisions from oracle-BLEU translations.

We use the decoder trained on these models to translate the training bitext. Along with the 1-best translation (based on model scores), we also store search graphs or lattices generated during the translations process. Using the target sentences, we convert the translation lattice to an isomorphic oracle-BLEU lattice which has the same set of nodes but the edges represent BLEU score differences corresponding to each transition. Finally, we extract n-best candidate translations from the graphs ranked on BLEU score as defined in Equation (3). Using the word alignments from the initial phrase table, we extract the alignments between each source sentence and each of their n-best oracle-BLEU translations. Finally, we re-train the phrase translations, re-ordering and BiLM on these translations and alignments.

3.3 Avoiding over-fitting

Re-estimation of the translation models from the n-best translation of the bitext could re-enforce the probabilities of the low frequency phrase pairs in the re-estimated models leading to over-fitting. Within forced decoding, Wuebker et al. (2010) address this problem by using a leave-one-out approach where they modify the phrase translation probabilities for each sentence pair by removing the counts of all phrases that were extracted from that particular sentence. However, in our approach, we do not impose a constraint to produce the exact translation, instead we use the highest BLEU translations which may be very different from the references. Thus it is not strictly necessary to apply leave-one-out in our approach as a solution to over-fitting. Instead, we handle the problem by simply removing all the phrase pairs below a threshold count which in our case is 2,

$$\phi_{init} = \phi_{baseline} - \phi_{C(e,f)<2} \quad (4)$$

therefore removing phrase pairs with high probability but low frequency.

4 Experimental set up

Our experiments are carried out for an Arabic-English parallel corpus of approximately 1 million sentence pairs. We establish a baseline system by training models on this bitext and then compare this to a forced decoding implementation and to oracle-BLEU re-estimation using the same bitext.

4.1 Baseline and forced decoding

The initial training corpus we use is a collection of parallel sentences taken from OpenMT data sources released by the LDC.

Phrase table, distortion models and the lexical BiLM are trained with initial alignments obtained using GIZA++ (Och and Ney, 2003). The English 5-gram target language model is trained with Kneser-Ney smoothing on news data of nearly 1.6B tokens. We use an in-house phrase-based SMT system similar to Moses. For all settings in this paper, weights were optimized on NIST's MT04 data set using pairwise ranked optimization (Hopkins and May, 2011).

For forced alignment we use the existing implementation within the Moses SMT toolkit (Koehn

Baseline	50.1		
	n=1	n=10	n=100
PT_{re}	50.1(0.0)	50.1(0.0)	50.0(-0.1)
PT_{in}	50.7$^{\blacktriangle}$(+0.6)	50.5$^{\blacktriangle}$(+0.4)	50.0(-0.1)
$BiLM_{re}$ + PT_{in}	50.9$^{\blacktriangle}$(+0.8)	50.5$^{\blacktriangle}$(+0.4)	49.6(-0.5)

Table 1: Performance of our oracle-BLEU re-estimation with varying size n of n-best lists for the MT09 test set. $^{\blacktriangle}/^{\blacktriangledown}$ indicates a statistically significant gain/drop at $p < 0.01$ and $^{\vartriangle}/^{\triangledown}$ at $p < 0.05$. Values in brackets show gains over the baseline.

et al., 2007) trained on the baseline phrase translation model. In order to increase the chances of producing the exact reference, we follow Foster and Kuhn (2012) and relax the standard decoding parameters as follows: distortion limit=∞, stack size=2000, beam width=10e-30, and no threshold pruning of the translation model.

4.2 Oracle BLEU re-estimation

To obtain oracle-BLEU translations, we first train an initial SMT system and use it to decode the bitext. This system is identical to the baseline system except for the removal of low-frequency phrase pairs from the baseline phrase table as described in Section 3.3. To obtain the n-best oracle-BLUE translations, we experiment with different values of n, where $n \in \{1, 10, 100\}$. From these oracle-BLEU translations and alignments all phrases that were used in the derivation of these n-best sentences are extracted and the models are re-estimated by re-calculating the translation probabilities. Hierarchical and lexicalized re-ordering models as well as the BiLM are re-trained using the source sentences, oracle-BLEU translations and word alignments. For testing the performance of the re-estimated models, we tune different systems while replacing the baseline models with the corresponding re-estimated models. We also experiment with the interpolation of re-estimated models with the respective baseline models. We evaluate against 4 test sets: MT05, MT06, MT08, and MT09. Case-insensitive 4-gram BLEU (Papineni et al., 2002) is used as evaluation metric. Approximate randomization (Noreen., 1989; Riezler and Maxwell, 2005) is used to detect statistically significant differences.

5 Results

We discuss the experimental results of our oracle-BLEU re-estimation approach for different models and settings and provide a comparison with the baseline (heuristic training) and forced alignment.

Re-estimated models with three different values of $n \in \{1, 10, 100\}$ were evaluated under three settings: phrase table re-estimation, interpolation, and BiLM re-estimation. The best improvements over the baseline are obtained by using only 1-best (n= 1) alignments as shown in Table 1. Surprisingly, this is in contrast with forced decoding as discussed in Wuebker et al. (2010), where the best improvements are obtained for $n = 100$.

Table 2 provides a comparison between BLEU improvements achieved by forced decoding ($n = 100$ best) and our oracle-BLEU re-estimation approach ($n = 1$ best) over the baseline for different models. One can see in Table 2 that while phrase table re-estimation drops substantially for forced decoding for all test sets (up to -1.4 for MT09), oracle-BLEU phrase table re-estimation shows either slight improvements or negligible drops compared to the baseline. For the linear interpolation of the re-estimated phrase table with the baseline, forced decoding shows only a slight improvement for MT06, MT08 and MT09 and still suffers from a substantial drop for MT05. On the other hand, oracle-BLEU re-estimation shows consistent improvements for all test sets with a maximum gain of up to +0.7 for MT06. It is important to note here that although linear interpolation extinguishes the advantage of a smaller phrase table size obtained by re-estimation, the improvement achieved by interpolation for oracle-BLEU re-estimation are significantly higher as compared to forced decoding.

An important novelty of oracle-BLEU re-estimation is that it also allows for re-training of other models alongside the phrase table. Here we provide the results for the re-estimation of a BiLM. For all test sets, BiLM re-estimation provides additional improvements over simple phrase table interpolation, demonstrating that re-estimation of re-ordering models can further improve translation performance. The last row of Table 2 shows that the re-estimated BiLM on its own adds BLEU improvement of up to +0.5 (for MT09). The highest BLEU improvement of +0.8 is achieved by using a re-estimated BiLM and an interpolated phrase table. Note that re-estimation of BiLM or re-ordering models is not possible for forced decoding due to the constraint of having to match the exact reference. For an additional anal-

	MT05		MT06		MT08		MT09	
Baseline	58.5		47.9		47.3		50.1	
	FD	OB	FD	OB	FD	OB	FD	OB
PT_{re}	57.4$^{\blacktriangledown}$(-1.1)	58.7$^{\triangle}$(+0.2)	46.3(-0.7)	47.8$^{\blacktriangledown}$(-0.1)	46.1$^{\blacktriangledown}$(-1.2)	47.4$^{\triangle}$(+0.1)	48.7$^{\blacktriangledown}$(-1.4)	50.1(0.0)
PT_{in}	58.2$^{\triangledown}$(-0.3)	58.8$^{\blacktriangle}$(+0.3)	48.0(+0.1)	**48.6**$^{\blacktriangle}$(+0.7)	47.5(+0.2)	47.7$^{\blacktriangle}$(+0.4)	50.4$^{\triangle}$(+0.3)	50.7$^{\blacktriangle}$(+0.6)
PT_{in}+ $BiLM_{re}$	-	**59.2**$^{\blacktriangle}$(+0.7)	-	48.5$^{\blacktriangle}$(+0.6)	-	**47.7**$^{\blacktriangle}$(+0.4)	-	**50.9**$^{\blacktriangle}$(+0.8)
PT_{base} + $BiLM_{re}$	-	58.6(+0.1)	-	48.2(+0.3)	-	47.2$^{\blacktriangledown}$(-0.1)	-	50.6(+0.5)

Table 2: BLEU scores for Forced decoding and Oracle BLEU re-estimation. $PT_{re/in}$ = Phrase table re-estimation/interpolation/baseline, PT_{base} = Baseline Phrase table, $BiLM_{re}$ = BiLM re-estimation, FD=Forced decoding, OB=oracle-BLEU.

	TEST	
Baseline	51.0	
	FD_{LO}	OB
PT_{re}	50.7$^{\triangledown}$(-0.3)	51.0 (0.0)
PT_{in}	51.5$^{\blacktriangle}$(+0.5)	51.5$^{\blacktriangle}$(+0.5)
PT_{in} + $BiLM_{re}$	-	**51.6**$^{\blacktriangle}$(+0.6)

Table 3: BLEU scores for Oracle-Bleu and Forced decoding with leave-one-out against concatenation of MT03, MT05-MT09.

	(% of baseline)
OB_{100}	5.07
OB_{10}	4.16
OB_1	**3.28**
FD	27.71
FD_{LO}	7.6

Table 4: Phrase table sizes compared to baseline for Oracle-BLUE re-estimation and Forced decoding for different n-best list sizes, FD_{LO} = Forced decoding with leave-one-out.

ysis, we experimented with the interpolation of both the re-estimated phrase table (forced decoding and oracle-BLEU) with the baseline. However, improvements achieved with this interpolation did not surpass the best result obtained for the oracle-BLEU re-estimation.

Additionally, we also compare oracle-BLEU re-estimation to forced decoding with leave-one-out (Wuebker et al., 2010) by evaluating both on a concatenation of 5 test sets (MT03, MT05-MT09). As shown in Table 3, even with leave-one-out, forced decoding performance drops below the baseline by -0.3 BLEU. In contrast, phrase tables re-estimated from oracle-BLEU translation achieves the same performance as the baseline. When interpolated with the baseline phrase table, both approaches show significant improvements over the baseline. This implies that only in combination with the original phrase table does forced-decoding with leave-one-out outperform the baseline. On the other hand, oracle-BLEU re-estimation by its own not only performs better than forced decoding, but also gives a performance equal to forced decoding with leave-one-out when interpolated with baseline phrase table. In addition to the BLEU improvements, our approach also results in a re-estimated phrase table with a significantly reduced size as compared to the baseline. As shown in Table 4, out of all the settings, the minimum phrase table size after oracle-BLEU re-estimation is only 3.28% of baseline (i.e., a reduction of 96.72%) while it is 7.6% for forced decoding.

6 Conclusions

In this paper, we proposed a novel technique for improving the reliability of SMT models by model re-estimation from oracle-BLEU translations of the source sentences in the bitext. Our experimental results show BLEU score improvements of up to +0.8 points for oracle-BLEU re-estimation over a strong baseline along with a substantially reduced size of the re-estimated phrase table (3.3% of the baseline). An important novelty of our approach is that it also allows for the re-estimation of re-ordering models which can yield further improvements in SMT performance as demonstrated by the re-estimation of a BiLM.

Acknowledgments

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project numbers 639.022.213 and 612.001.218. We thank Arianna Bisazza and the anonymous reviewers for their comments.

References

David Chiang, Yuval Marton, and Philip Resnik. 2008. Online large-margin training of syntactic and structural translation features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Markus Dreyer, Keith Hall, and Sanjeev Khudanpur. 2007. Comparing reordering constraints for smt using efficient BLEU oracle computation. In *Proceedings of 2007 Workshop on Syntax and Structure in Statistical Trans- lation*.

George Foster and Roland Kuhn. 2012. Forced decoding for phrase extraction. Technical report, University of Montreal.

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*. Association for Computational Linguistics.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ekaterina Garmash and Christof Monz. 2014. Dependency-based bilingual language models for reordering in statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1689–1700.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. 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, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 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*. Association for Computational Linguistics.

Patrik Lambert, Holger Schwenk, Christophe Servan, and Sadaf Abdul-Rauf. 2011. Investigations on translation model adaptation using monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics.

Percy Liang, Alexandre Bouchard-Côté, Dan Klein, and Ben Taskar. 2006. An end-to-end discriminative approach to machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July.

Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jan Niehues, Teresa Herrmann, Stephan Vogel, and Alex Waibel. 2011. Wider context by using bilingual language models in machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics.

Eric W. Noreen. 1989. *Computer Intensive Methods for Testing Hypotheses. An Introduction*. Wiley-Interscience.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Holger Schwenk, Patrik Lambert, Loïc Barrault, Christophe Servan, Haithem Afli, Sadaf Abdul-Rauf, and Kashif Shah. 2011. LIUM's SMT machine translation systems for WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics.

Ankit K. Srivastava, Yanjun Ma, and Andy Way. 2011. Oracle-based training for phrase-based statistical machine translation. In *Proceedings of the 15th annual meeting of the European Association for Machine Translation*.

Christoph Tillman. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL*. Association for Computational Linguistics.

Taro Watanabe, Jun Suzuki, Hajime Tsukada, and Hideki Isozaki. 2007. Online large-margin training for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational*

Natural Language Learning (EMNLP-CoNLL). Association for Computational Linguistics.

Guillaume Wisniewski, Alexandre Allauzen, and François Yvon. 2010. Assessing phrase-based translation models with oracle decoding. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Joern Wuebker, Arne Mauser, and Hermann Ney. 2010. Training phrase translation models with leaving-one-out. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July.