Towards better Machine Translation Quality for the German–English Language Pairs

Philipp Koehn Abhishek Arun Hieu Hoang
School of Informatics
University of Edinburgh
pkoehn@inf.ed.ac.uk a.arun@sms.ed.ac.uk h.hoang@sms.ed.ac.uk

Abstract

The Edinburgh submissions to the shared task of the Third Workshop on Statistical Machine Translation (WMT-2008) incorporate recent advances to the open source Moses system. We made a special effort on the German–English and English–German language pairs, leading to substantial improvements.

1 Introduction

Edinburgh University participated in the shared task of the Third Workshop on Statistical Machine Translation (WMT-2008), which is partly funded by the EUROMATRIX project, which also funds our work. In this project, we set out to build machine translation systems for all language pairs of official EU languages. Hence, we also participated in the shared task in all language pairs.

For all language pairs, we used the Moses decoder (Koehn et al., 2007), which follows the phrase-based statistical machine translation approach (Koehn et al., 2003), with default settings as a starting point. We recently added minimum Bayes risk decoding and reordering constraints to the decoder. We achieved consistent increase in BLEU scores with these improvements, showing gains of up to 0.9% BLEU on the 2008 news test set.

Most of our efforts were focused on the language pairs German–English and English–German. For both language pairs, we explored language-specific and more general improvements, resulting in gains of up to 1.5% BLEU for German–English and 1.4% BLEU for English–German.

2 Recent Improvements

Over the last months, we added minimum Bayes risk decoding and additional reordering constraints to the Moses decoder. The WMT-2008 shared task offered the opportunity to assess these components over a large range of language pairs and tasks.

For all our experiments, we trained solely on the Europarl corpus, which allowed us to treat the 2007 news commentary test set (nc-test2007) as a stand-in for the 2008 news test set (news-2008), for which we have no in-domain training data. This may have resulted in lower performance due to less (and very relevant) training data, but it also allowed us to optimize for a true out-of-domain test set.

The baseline training uses Moses default parameters. We use a maximum sentence length of 80, a phrase translation table with the five traditional features, lexicalized reordering, and lowercase training and test data. All reported BLEU scores are not case-sensitive, computed using the NIST tool.

2.1 Minimum Bayes Risk Decoding

Minimum Bayes risk decoding was proposed by Kumar and Byrne (2004). Instead of selecting the translation with the highest probability, minimum Bayes risk decoding selects the translation that is most similar to the highest scoring translations. Intuitively, this avoid the selection of an outlier as the best translation, since the decision rule prefers translations that are similar to other high-scoring translations.

Minimum Bayes risk decoding is defined as:

$$\mathbf{e}_{\text{MBR}} = \text{argmax}_{\mathbf{e}} \sum_{\mathbf{e}'} L(\mathbf{e}, \mathbf{e}')\ p(\mathbf{e}'|\mathbf{f})$$

As similarity function L, we use sentence-level BLEU with add-one smoothing. As highest scoring translations, we consider the top 100 distinct translations, for which we convert the translation scores into a probability distribution p (with a scaling factor of 1). We tried other n-best list sizes and scaling factors, with very similar outcomes.

Proceedings of the Third Workshop on Statistical Machine Translation, pages 139–142,
Columbus, June 2008.

Language Pair	Baseline	MBR	MP	MBR+MP
Spanish–German news	11.7	11.8 (+0.1)	11.9 (+0.2)	12.0 (+0.3)
Spanish–German ep	20.7	21.0 (+0.3)	20.8 (+0.1)	21.0 (+0.3)
German–Spanish news	16.2	16.3 (+0.1)	16.4 (+0.2)	16.6 (+0.4)
German–Spanish ep	28.5	28.6 (+0.1)	28.5 (±0.0)	28.6 (+0.1)
Spanish–English news	19.8	20.2 (+0.4)	20.2 (+0.4)	20.3 (+0.5)
Spanish–English ep	33.6	33.7 (+0.1)	33.6 (±0.0)	33.7 (+0.1)
English–Spanish news	20.1	20.5 (+0.4)	20.5 (+0.4)	20.7 (+0.6)
English–Spanish ep	33.1	33.1 (±0.0)	33.0 (–0.1)	33.1 (±0.0)
French–English news	18.5	19.1 (+0.6)	19.1 (+0.6)	19.2 (+0.7)
French–English ep	33.5	33.5 (±0.0)	33.4 (–0.1)	33.5 (±0.0)
English–French news	17.8	18.0 (+0.2)	18.2 (+0.4)	18.3 (+0.5)
English–French ep	31.1	31.1 (±0.0)	31.1 (±0.0)	31.1 (±0.0)
Czech–English news	14.2	14.4 (+0.2)	14.3 (+0.1)	14.5 (+0.3)
Czech–English nc	22.8	23.0 (+0.2)	22.9 (+0.2)	23.0 (+0.2)
English–Czech news	9.6	9.6 (±0.0)	9.7 (+0.1)	9.6 (±0.0)
English–Czech nc	12.9	13.0 (+0.1)	12.9 (±0.0)	13.0 (+0.1)
Hungarian–English news	7.9	8.3 (+0.4)	8.5 (+0.6)	8.8 (+0.9)
English–Hungarian news	6.1	6.3 (+0.2)	6.4 (+0.3)	6.5 (+0.4)
average news	-	+0.26	+0.33	+0.46
average ep	-	+0.08	–0.02	+0.08

Table 1: Improvements in BLEU on the test sets test2008 (ep), newstest2008 (news) and nc-test2008 (nc) for minimum Bayes risk decoding (MBR) and the monotone-at-punctuation reordering (MP) constraint.

2.2 Monotone at Punctuation

The reordering models in phrase-based translation systems are known to be weak, since they essentially relies on the interplay of language model, a general preference for monotone translation, and (in the case of lexicalized reordering) a local model based on a window of neighboring phrase translations. Allowing any kind of reordering typically reduces translation performance, so reordering is limited to a window of (in our case) six words.

One noticeable weakness is that the current model frequently reorders words beyond clause boundaries, which is almost never well-motivated, and leads to confusing translations. Since clause boundaries are often indicated by punctuation such as comma, colon, or semicolon, it is straight-forward to introduce a reordering constraint that addresses this problem.

Our implementation of a monotone-at-punctuation reordering constraint (Tillmann and Ney, 2003) requires that all input words before clause-separating punctuation have be translated, before words afterwards are covered. Note that this constraint does not limit in any way phrase translations that span punctuation.

2.3 Results

Table 1 summarizes the impact of minimum Bayes risk decoding (MBR) and the monotone-at-punctuation reordering constraint (MP). Scores show higher gains for out-of-domain news test sets (+0.46) than for in-domain Europarl sets (+0.08).

3 German–English

Translating between German and English is surprisingly difficult, given that the languages are closely related. The main sources for this difficulty is the different syntactic structure at the clause level and the rich German morphology, including the merging of noun compounds.

In prior work, we addressed **reordering** with a pre-order model that transforms German for training and testing according to a set of hand-crafted rules (Collins et al., 2005). Employing this method to our baseline system leads to an improvement of +0.8 BLEU on the nc-test2007 set and +0.5 BLEU on the test2007 set.

German–English	nc-test2007	test2007
baseline	20.3	27.6
tokenize hyphens	20.1 (–0.2)	27.6 (±0.0)
tok. hyph. + truecase	20.7 (+0.4)	27.8 (+0.2)

Table 2: Impact of truecasing on case-sensitive BLEU

In a more integrated approach, factored translation models (Koehn and Hoang, 2007) allow us to consider grammatical coherence in form of **part-of-speech language models**. When translating into output words, we also generate a part-of-speech tag along with each output word. Since there are only 46 POS tags in English, we are able to train high-order n-gram models of these sequences. In our experiments, we used a 7-gram model, yielding improvements of +0.2/–0.1. We obtained the POS tags using Brill's tagger (Brill, 1995).

Next, we considered the problem of unknown input words, which is partly due to hyphenated words, noun compounds, and morphological variants. Using the baseline model, 907 words (1.78%) in nc-test2007 and 262 (0.47%) in test2007 are unknown. First we separate our **hyphens** by tokenizing words such as *high-risk* into *high @-@ risk*. This reduces the number of unknown words to 791/224. Unfortunately, it hurts us in terms of BLEU (–0.1/–0.1). Second, we **split compounds** using the frequency-based method (Koehn and Knight, 2003), reducing the number of unknown words to than half, 424/94, improving BLEU on nc-test2007 (+0.5/–0.2).

A final modification to the data preparation is **truecasing**. Traditionally, we lowercase all training and test data, but especially in German, case marks important distinctions. German nouns are capitalized, and keeping case allows us to make the distinction between, say, the noun *Wissen* (*knowledge*) and the verb *wissen* (*to know*). By truecasing, we only change the case of the first word of a sentence to its most common form. This method still needs some refinements, such as the handling of headlines or all-caps text, but it did improve performance over the hyphen-tokenized baseline (+0.3/+0.2) and the original baseline (+0.2/+0.1).

Note that truecasing simplifies the recasing problem, so a better way to gauge its effect is to look at the case-sensitive BLEU score. Here the difference are slightly larger over both the hyphen-tokenized baseline (+0.6/+0.2) and the original baseline (+0.4/+0.2). See the Table 2 for details.

German–English	nc-test2007	test2007
baseline	21.3	28.4
pos lm	21.5 (+0.2)	28.3 (–0.1)
reorder	22.1 (+0.8)	28.9 (+0.5)
tokenize hyphens	21.2 (–0.1)	28.3 (–0.1)
tok. hyph. + split	21.8 (+0.5)	28.2 (–0.2)
tok. hyph. + truecase	21.5 (+0.2)	28.5 (+0.1)
mp	21.6 (+0.3)	28.2 (–0.2)
mbr	21.4 (+0.1)	28.3 (–0.1)
big beam	21.3 (±0.0)	28.3 (–0.1)

Table 3: Impact of individual modifications for German–English, measured in BLEU on the development sets

German–English	nc-test2007	test2007
baseline	21.3	28.4
+ reorder	22.1 (+0.8)	28.9 (+0.5)
+ tokenize hyphens	22.1 (+0.8)	28.9 (+0.5)
+ truecase	22.7 (+1.3)	28.9 (+0.5)
+ split	23.0 (+1.7)	29.1 (+0.7)
+ mbr	23.1 (+1.8)	29.3 (+0.9)
+ mp	23.3 (+2.0)	29.2 (+0.8)

Table 4: Impact of combined modifications for German–English, measured in BLEU on the development sets

As for the other language pairs, using the **monotone-at-punctuation** reordering constraint (+0.3/–0.2) and **minimum Bayes risk decoding** (+0.1/–0.1) mostly helps. We also tried **bigger beam** sizes (stack size 1000, phrase table limit 50), but without gains in BLEU (±0.0/–0.1).

Table 3 summarizes the contributions of the individual modifications we described above. For our final system, we added the improvements one by one (see Table 4), except for the bigger beam size and the POS language model. This led to an overall increase of +2.0/+0.8 over the baseline. Due to a bug in splitting, the system we submitted to the shared task had a score of only +1.5/+0.6 over the baseline.

4 English–German

For English–German, we applied many of the same methods as for the inverse language pair. Tokenizing out **hyphens** has questionable impact (–0.1/+0.1), while **truecasing** shows minor gains (±0.0/+0.1), slightly higher for case-sensitive scoring (+0.2/+0.3). We have not yet developed a method that is the analog of the compound splitting

English–German	nc-test2007	test-2007
baseline	14.6	21.0
tokenize hyphens	14.5 (–0.1)	21.1 (+0.1)
tok. hyph. + truecase	14.6 (±0.0)	21.1 (+0.1)
morph lm	15.7 (+1.1)	21.2 (+0.2)
mbr	14.9 (+0.3)	21.0 (±0.0)
mp	14.8 (+0.2)	20.9 (–0.1)
big beam	14.7 (+0.1)	21.0 (±0.0)

Table 5: Impact of individual modifications for English–German, measured in BLEU on the development sets

method — compound merging. We consider this an interesting challenge for future work.

While the rich German morphology on the source side mostly poses sparse data problems, on the target side it creates the problem of which morphological variant to choose. The right selection hinges on grammatical agreement within noun phrases, the role that each noun phrase plays in the clause, and the grammatical nature of the subject of a verb. We use LoPar (Schmidt and Schulte im Walde, 2000), which gives us **morphological features** such as case, gender, count, although in limited form, it often opts for more general categories such as *not genitive*. We include these features in a sequence model, as we used a sequence model over part-of-speech tags previously. The gains of this method are especially strong for the out-of-domain set (+1.1/+0.2).

Minimum Bayes risk decoding (+0.3/±0.0), the **monotone-at-punctuation** reordering constraint (+0.2/–0.1), and **bigger beam sizes** (+0.1/±0.0) have similar impact as for the other language pairs. See Table 5 for a summary of all modifications. By combining everything except for the bigger beam size, we obtain overall gains of +1.4/+0.4 over the baseline. For details, refer to Table 6.

5 Conclusions

We built Moses systems trained on either only Europarl data or, for Czech and Hungarian, the available training data. We showed gains with minimum Bayes risk decoding and a reordering constraint involving punctuation. For German↔English, we employed further language-specific improvements.

Acknowledgements: This work was supported in part under the EuroMatrix project funded by the European Commission (6th Framework Programme).

English–German	nc-test2007	test2007
baseline	14.6	21.0
+ tokenize hyphens	14.5 (–0.1)	21.1 (+0.1)
+ truecase	14.6 (±0.0)	21.1 (+0.1)
+ morph lm	15.4 (+0.8)	21.3 (+0.3)
+ mbr	15.7 (+1.1)	21.4 (+0.4)
+ mp	16.0 (+1.4)	21.4 (+0.4)

Table 6: Impact of combined modifications for English–German, measured in BLEU on the development sets

References

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*, 21(4).

Collins, M., Koehn, P., and Kucerova, I. (2005). Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531–540, Ann Arbor, Michigan. Association for Computational Linguistics.

Koehn, P. and Hoang, H. (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.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). 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. Association for Computational Linguistics.

Koehn, P. and Knight, K. (2003). Empirical methods for compound splitting. In *Proceedings of Meeting of the European Chapter of the Association of Computational Linguistics (EACL)*.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*.

Kumar, S. and Byrne, W. (2004). Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*.

Schmidt, H. and Schulte im Walde, S. (2000). Robust German noun chunking with a probabilistic context-free grammar. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Tillmann, C. and Ney, H. (2003). Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational Linguistics*, 29(1).