Edinburgh University System Description for the 2008 NIST Machine Translation Evaluation

Philipp Koehn, Josh Schroeder and Miles Osborne
School of Informatics
University of Edinburgh

Abstract

The MT Group at Edinburgh University participated in the constraint task track of the 2008 NIST MT evaluation for three language pairs: Arabic–English, Chinese–English, and Urdu–English. We built systems using our Moses decoder, with MBR decoding, and large language models within the limits of 32-bit machines.

1 Common Setup

The basic setup is the Moses decoder (Koehn et al., 2007), which is publicly available at *http://www.statmt.org/*. The Moses system allows the training, tuning, and testing of statistical machine translation systems, when provided with the parallel corpora such as the ones made available for the NIST evaluation campaign.

From this starting point, the main challenge of the NIST evaluation is to make use of the large training corpora available for the language pairs Arabic–English and Chinese–English. We were constraint in our experimentation by our 32-bit infrastructure, which limits process size to 3 GB.

One key advantage of Moses is the use of on-disk translation models (Zens and Ney, 2007), which leaves the available RAM for the language model. Nevertheless, we are limited to 4-gram models trained on one billion words, while much more language modelling data is available. We are aware that better performance is possible (Brants et al., 2007), if more memory would be available, or more efficient use of the available memory were made.

While there are gains possible with language specific methods, we did not make use of anything besides a Chinese number translator. We note that gains are possible using Chinese sentence restructuring (Wang et al., 2007) and basic Arabic morphological preprocessing.

The training regime our systems can be summarized as follows:

- sentence length limit 80 words
- GIZA++ training
- word alignment heuristic *grow-diag-final-and*
- phrase length limit 7
- SRILM toolkit with interpolated Kneser-Ney discounting
- three separate language models trained on
  - English side of parallel corpus
  - AFP part of Gigaword corpus
  - Xinhua part of Gigaword corpus
- lexicalized reordering model with option *msd-bidirectional-fe*
- recaser trained as monotone translation model
- weights optimized with max BLEU training

We spent about two months (part time) on getting our systems in shape for the 2008 NIST evaluation. Note that training large systems easily takes 1–2 weeks, mostly due to the slow GIZA++ word alignment stage.

While the engineering of a system for such an evaluation mostly involves adapting existing methods on the task at hand, and not the development of new methods, it nevertheless is a crucial stress test and helped us to track down some bugs, most notably with the broken MBR decoding and some language modelling issues. All these improvements are included in the last Moses release.

However, some of our latest advances, especially the use of factored translation models (Koehn and Hoang, 2007), randomized language models (Talbot and Osborne, 2007), domain adaptation methods (Koehn and Schroeder, 2007), and a better recaser did not make it into the final systems due the limited time for experimentation.

2 Experiments for Chinese–English

We performed a number of experiments to find the optimal setup for our Chinese–English system. As development test set, we used the evaluation set from the 2006 NIST evaluation. Automatic parameter tuning was carried out on the 2002 NIST evaluation set.

A notable property of the training of Chinese–English systems with the available training data is that a limited selection of the training data (excluding UN and Hongkong data) with more heavily *news* domain data leads to better performance. Part of the reason may be that we are able to train 5-gram language models on the limited news data, but only 3-gram language models on all the data.

Chinese–English	BLEU
news (5g lm)	33.62
all (3g lm)	31.95

Note that all reported BLEU scores are non-case sensitive scores.

Better performance may be achieved when using a larger beam (translation table limit 50 and stack size limit 1000), although this does slow down decoding drastically.

Chinese–English	BLEU
news (5g lm)	33.62
bigger beam	33.94

For the following set of experiments, we fixed a bug with regard to language modelling, otherwise the same setup is used as in the above news 5gram system. We achieved additional gains using minimum Bayes risk decoding (Kumar and Byrne, 2004), using the evaluation sets from 2002 to 2005 as training data, and rules for Chinese numbers and tag-lines.

Chinese–English	BLEU
baseline	34.85
baseline + MBR	35.28
baseline + eval 02-05	35.38
baseline + rules	35.63

In the final setup, we used all these three improvements, as well as order 4 instead of 3 for the two Gigaword language models.

Chinese–English	BLEU
final	36.81

3 Experiments for Arabic–English

We performed a number of experiments to find the optimal setup for our Arabic–English system. As development test set, we used the evaluation set from the 2006 NIST evaluation. Automatic parameter tuning was carried out on the 2004 NIST evaluation set.

Opposed to the Chinese–English experience, using all the training data for Arabic helps.

Arabic–English	BLEU
news (3g lm)	38.12
all (3g lm)	40.42

The runs above do not include the Gigaword language models, which lead to better performance. Even more so, when 4-gram language models are used. The best language model setup was the use of a single interpolated 4-gram language model with interpolation weights tuned on the NIST 2004 evaluation set.

Arabic–English	BLEU
baseline	40.42
baseline + Gigaword, 3g lm	41.01
baseline + Gigaword, 4g lm	42.08
baseline + interpolated lm	43.06

Final improvements were the use of rules for tag line translations and MBR decoding.

Arabic–English	BLEU
baseline	43.06
+ tag line rules	43.89
+ MBR	43.94

4 Experiments for Urdu–English

For Urdu–English, we were limited to the provided training data, which meant that training was much quicker, and there were also no memory issues with large language models. The challenges with this task were instead dealing with the sparse training data.

To create reliable word alignments, we experimented with stemming the Urdu input. Stemming helped with creating less sparse source data. When we were training on shorter sentences and only the ”translations” portion of the training data, stemming during word alignment helped our BLEU test scores. When we expanded the training data to include the

"found" and "special" data sets, and increased the maximum sentence length, the benefits of stemming for word alignments diminished.

Unlike with the Chinese and Arabic tasks, we did not see gains with MBR decoding and larger beam sizes in the Urdu task.

We also experimented with the Berkeley word alignment package, which implements a joint training model with posterior decoding as described by Liang et al. (2006). We found the Berkeley aligner to be more robust than using GIZA++ in situations where we had long sentences and sparse word counts. In some of these more difficult training scenarios GIZA++ would crash or provide defective alignments. However, when GIZA++ was able to cope with the input and successfully trained, the phrase table generated from its word alignments produced slightly higher scoring translations than phrase tables created with the output of the Berkeley aligner.

The Berkeley aligner features a number of parameters, such as the posterior decoding threshold, that can be tuned to change final word alignments. Liang et al. (2006) tune these parameters to reduce AER. Instead of focusing on word alignments and AER, we experimented with various settings of the the posterior decoding threshold to improve BLEU score. While our final submission used one of the successful GIZA++ runs, we are actively investigating the best way to tune the Berkeley aligner parameters with an eye towards improving a final translation score rather than word alignment accuracy.

5 Acknowledgement

This work was supported in part under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022.

References

Brants, T., Popat, A. C., Xu, P., Och, F. J., and Dean, J. (2007). Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867.

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 Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, Prague, Czech Republic. Association for Computational Linguistics.

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)*.

Liang, P., Taskar, B., and Klein, D. (2006). Alignment by agreement. In *Human Language Technology and North American Association for Computational Linguistics (HLT/NAACL)*.

Talbot, D. and Osborne, M. (2007). Smoothed Bloom filter language models: Tera-scale LMs on the cheap. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 468–476.

Wang, C., Collins, M., and Koehn, P. (2007). Chinese syntactic reordering 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)*, pages 737–745.

Zens, R. and Ney, H. (2007). Efficient phrase-table representation for machine translation with applications to online MT and speech translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 492–499, Rochester, New York. Association for Computational Linguistics.