Edinburgh System Description for the 2005 NIST MT Evaluation

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, David Talbot, Michael White
School of Informatics
University of Edinburgh

Abstract

This document describes the first NIST MT Evaluation submission of the newly formed Edinburgh University Statistical Machine Translation Group. Our entry to the 2005 DARPA/NIST MT Evaluation was largely based on the 2004 MIT system. In a two month effort we focused on adding more data and a few new features to our Arabic-English system. We also worked on preprocessing and applied some of the lessons learnt to our Chinese-English system. Our efforts resulted in improved translation performance over the previous 2004 system. Competing in the competition was also a valuable learning experience in large-scale system building.

This document describes the first NIST MT Evaluation submission from Edinburgh University's Statistical Machine Translation Group. Our entry to this year's evaluation was based on the 2004 system from MIT (Koehn, 2004a). This was built by Philipp Koehn, who is now a faculty member in Edinburgh's School of Informatics. In a two month effort, the group familiarised itself with the system and the data, and added a few new features to the Arabic-English system.

1 Baseline System

The baseline system which we started with was a phrase-based statistical machine translation that used the Pharaoh decoder (Koehn, 2004b). As is common in current state-of-the-art statistical machine translation systems, we employed a log linear approach in our translation system. We searched for the most probable English sentence $\mathbf{e}$ given some foreign sentence $\mathbf{f}$ by maximising the sum over a set of feature functions $h_m(\mathbf{e}, \mathbf{f})$:

$$\hat{\mathbf{e}} = \arg\max_e p(\mathbf{e}|\mathbf{f}) \quad (1)$$

$$= \arg\max_e \sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}, \mathbf{f}) \quad (2)$$

A number of feature functions were employed when scoring candidate translation:

- Language model
- Phrase translation probability (both directions)
- Lexical translation probability (both directions)
- Word penalty
- Phrase penalty
- Linear reordering penalty

The language model was a smoothed trigram model created and trained using the English side of the Arabic parallel corpus (Stolke, 2002).

The phrase translation probability is defined as

$$p(\bar{f}|\bar{e}) = \frac{count(\bar{f}, \bar{e})}{\sum_{\bar{f}} count(\bar{f}, \bar{e})} \quad (3)$$

where $count(\bar{f}, \bar{e})$ gives the total number of times the phrase $\bar{f}$ was aligned with the phrase $\bar{e}$ in the parallel corpus.

Phrase translation probabilities are lexically weighted as in (Koehn et al., 2003):

$$p_{lw}(\bar{f}|\bar{e}, \mathbf{a}) = \prod_{i=1}^{n} \frac{1}{|\{i|(i,j) \in \mathbf{a}\}|} \sum_{\forall(i,j)\in\mathbf{a}} p(f_j|e_i) \quad (4)$$

where n is the length of $\bar{e}$, and $\mathbf{a}$ is the word-level alignment between phrase $\bar{e}$ and $\bar{f}$. Since a phrase alignment $< \bar{f}, \bar{e} >$ may have multiple possible word-level alignments, we retain a set of alignments and take the most frequent.

Word and phrase penalty add a constant factor (ω and π) for each word or phrase generated. The reordering penalty adds a factor δ^n for movements over n words.

The weight of these feature functions is set by minimum error rate training (Och, 2003). We thank David Chiang of the University of Maryland for providing us with a faster version of our implementation.

2 Improvements

Given the short time frame and the fact that this was our first group entry, we concentrated our efforts of implementing ideas that seem to help the other systems in the 2004 evaluation.

The 2004 system was trained on news data and on half of the UN corpus. Training the system on the additional half of the UN corpus gave us an improvement of (absolute) 2% BLEU of the development set (most of the 2002 evaluation set).

We added part of the English Gigaword Corpus to the training data for the language model. Training a language model on all this data exceeded the ability of our computing resources (32-bit, 5GB multiprocessor Linux machines). The largest language model we were able to train used a 800 million words, with all digits replaced by a single digit, and with all singleton trigrams pruned. This larger language model (compared with the one used for the 2004 entry) improved the system score by 2% BLEU on the development set.

We picked up an additional 2% BLEU improvement with a few additional system changes:

- Dropping unknown words during decoding
- Delete word feature
- Limited changes to the recapitaliser
- Limited post-editing of the output (largely changing UK-style dates to US-style dates)
- Limited changes to the tokenisation of Arabic

A number of other efforts were too premature at the time of the evaluation to yield system improvements:

- Better tokenisation Arabic and English
- Better recapitaliser
- Better reordering model
- Domain marking features
- Stemming for better word alignment
- Using POS tags for language modeling

Due to an oversight, we did not receive the expanded set of training data available for the 2005 evaluation competition, and instead had to rely on the training data used in the 2004 system. This meant, for example, that some names were untranslatable as they were present in the test set but not in the older training set. We expect that further improvements would be had by incorporating this additional training data, and plan to do a post hoc analysis using this extra data.

3 Results

Our system improved by (absolute) 6% on our development set – the second half of the 2002 Evaluation set. Gains on the other test sets for Arabic–English translation were as follows, as measured with %BLEU (Papineni et al., 2002):

Arabic–English	**'04 system**	**'05 system**
Eval 2002 (partial)	34.4	40.4
Eval 2004	34.1	34.3
Eval 2005	35.6	40.5

We noted that tuning on 300 sentences of the 2002 evaluation set cause too short output for the 2004

evaluation set – the 34.3% BLEU score reflects a 9.5% length penalty. With a length penalty optimised on the 2004 evaluation set (note: optimised on test), we could obtain a score of 37.7% BLEU on this set.

Our best system (submitted as contrastive run, with a score of 40.5% BLEU) was optimised on the first 500 sentences of the 2004 evaluation set, and also includes a specialised news language model. Our primary submission was optimised on the 2002 evaluation set, it scored 39.7% BLEU.

For the Chinese–English system we added a larger language model, and improved number translation. Otherwise it is identical to the 2004 system.

Chinese–English	**'04 system**	**'05 system**
Eval 2002 (partial)	26.1	27.2
Eval 2004	27.1	28.1
Eval 2005	24.3	25.1

References

Koehn, P. (2004a). The foundation for statistical machine translation at MIT. In *Proceedings of Machine Translation Evaluation Workshop 2004*.

Koehn, P. (2004b). Pharaoh: a beam search decoder for statistical machine translation. In *6th Conference of the Association for Machine Translation in the Americas, AMTA*, Lecture Notes in Computer Science. Springer.

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)*.

Och, F. J. (2003). Minimum error rate training for statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics (ACL)*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL)*.

Stolke, A. (2002). SRILM - an extensible language modeling toolkit. In *Proceedings of the International Confonference on Spoken Language Processing*.