HKUST Statistical Machine Translation Experiments for CWMT 2009

Chi-kiu LO, Dekai WU
Human Language Technology Center
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
E-mail: {jackielo, dekai}@cs.ust.hk

Abstract: This paper describes the HKUST experiments in the CWMT 2009 evaluation campaign on machine translation. We report results on the four tasks we joined, which are the Chinese to English single system translation in the news area, English to Chinese machine translation in the news area and the science and technology area and Chinese to Mongolian machine translation on daily expressions.

香港科技大学 CWMT2009 机器翻译评测系统

罗致翘 吴德恺
香港科技大学 计算机科学及工程系 人类语言技术中心
E-mail: {jackielo, dekai}@cs.ust.hk

摘 要:本文叙述了在 CWMT2009 机器翻译评测中香港科技大学的参评系统。我们参与了其中四个评测项目，当中包括汉英新闻领域单一系统，英汉新闻领域机器翻译，英汉科技领域机器翻译和汉蒙日常用语机器翻译。我们汇报我们的系统在这四个项目中的评测结果。

1 Introduction

We describe experiments conducted at HKUST during the CWMT 2009 evaluation campaign on machine translation. For our first participation in the CWMT evaluation, we aimed to understand the training and testing data used and the evaluation standard in CWMT 2009. We joined four tasks which included Chinese to English single system translation in the news area, English to Chinese machine translation in the news area and the science and technology area and Chinese to Mongolian machine translation on daily expressions. All the participating systems were pure phrase-based Statistical Machine Translation (SMT). Our systems were trained only on the provided training data using the publicly available toolkits with all the off-the-shelf settings. We deliberately excluded all external resources, such as the GALE data or any other lexicon dictionary, in all training steps.

2 System

2.1 Phrasal bi-lexicon

The phrasal bi-lexicon is obtained by extracting phrase pairs that are consistent with the IBM model 4 word alignments obtained with bidirectional GIZA++ (Och and Ney, 2002).

We used the Pharoah training script (Koehn, 2004) with the grow-diag-final heuristic in phrasal extraction. The grow-diag-final heuristic expands the word alignment by adding directly neighboring alignment points, and alignment points in the diagonal neighborhood. We trained relatively long phrasal translations, allowing phrases of length up to 20 words.

2.2 Language Model

The language models are trained with Kneser-Ney smoothing using the SRI language modeling toolkit (Stolcke, 2002). For all the evaluation tasks, the primary language model was a 6-gram model trained on the target language of the bi-lexicon training data. For the Chinese to Mongolian machine translation task, we used an additional 6-gram language model trained on the monolingual Mongolian training corpus.

2.3 Decoder

We used Moses decoder (Koehn et al., 2007), which is an open source toolkit for statistical machine translation. Moses uses a log-linear model, which combines several knowledge sources in translation decision. It is a factored phrase-based beam-search decoder which represents each input word as a factor rather than the word surface form only. A factored decoder allows the translation model incorporate, in addition to the surface forms, richer linguistic information, such as part-of-speech, class and morphology. However, we did not use the factored representation in our evaluation experiments. We used the surface form of words only. We found that Moses decoder achieves slightly higher performance than its close-source predecessor Pharaoh in previous study (Shen et al., 2007).

3 Data

Table 1: Training data statistics computed for the phrasal bi-lexicon of the 4 evaluation tasks

Training Data Statistics	**Zh-En News**	**En-Zh News**	**En-Zh Sci-tech**	**Zh-Mn Daily**
Number of bi-sentences	1362848	1362848	925750	67251
Vocabulary size (input lang)	30560735	33749326	25080053	812753
Vocabulary size (output lang)	33749326	30560735	22731434	811229

Table 2: Training data statistics computed for the additional language model of Chinese-Mongolian daily expression machine translation task

Number of sentences (Mongolian)	62399
Vocabulary size (Mongolian)	998629

3.1 Data description

The training set for the phrasal bi-lexicon of each evaluation task was drawn from the resources provided by the organizer. The resources were examined manually in files level and only files in the same domain as the evaluation task were used for training. Table 1 shows the training data statistics for the phrasal bi-lexicon of each of the evaluation task. The Chinese to English single system translation in news area and the English to Chinese machine translation in news area tasks were trained on the same set of data with the source and target language interchanged.

An additional 6-gram language model was trained for the Chinese to Mongolian machine translation task on Mongolian monolingual training data. Table 2 shows the training data statistics for the additional language model of the Chinese to Mongolian machine translation on daily expressions task.

3.2 Training data preprocessing

Before the bi-sentences with fertility ratio greater than 8 were filtered out in the last step before training. The training data was preprocessed with a language-specific but simple scheme for tokenization and normalization.

3.2.1 Chinese

We used the HKUST maximum entropy Chinese parser to tokenize the Chinese side of the corpus. A number segmenter was applied to fix the wrongly segmented number or time

expression.

3.2.2 English

The English sentences of the training corpus were tokenized and case-normalized. Case-normalization was done by normalizing the first word of the sentence to its most frequent form. (Zollmann et.al., 2006) A number segmenter was applied to fix the wrongly segmented number or time expression.

3.2.3 Mongolian

We only performed basic tokenization for Mongolian. No language specific preprocessing was done on the Mongolian side of corpus.

4 Experiments

4.1 Testing input processing

For the testing data, we used the same basic preprocessing as the training data. The testing input was preprocessed with a language-specific but simple scheme for tokenization and normalization.

4.1.1 Chinese processing in Chinese to English single system translation in news area

We used the HKUST maximum entropy Chinese parser to segment the Chinese testing input in the Chinese to English single system translation in news area. Number expression marker was then applied to markup the number and time expression and to provide alternative phrasal translation for the decoder to consider during translation.

4.1.2 English processing in the two English to Chinese machine translation tasks

The English testing input in the two English to Chinese machine translation tasks were tokenized and case-normalized. Case-normalization was done by normalizing the first word of the sentence.

4.1.3 Chinese processing in Chinese to Mongolian machine translation on daily expressions task

The Chinese testing input in the Chinese to Mongolian machine translation on daily expressions task was segmented by the HKUST maximum entropy Chinese parser. No further expression marker was applied in this task.

4.2 Experiment setup

For all the evaluation tasks, with the basic objective to understand the training and testing data, we did not do any tuning or optimization. We run the experiments with the default parameters and weights. The additional language model in the Chinese to Mongolian machine translation on daily expressions task was weighted the same as the primary language model.

4.3 Translation output processing

For all the evaluation tasks, we used simple heuristics to clean and normalize punctuation, capitalization and contractions in the translation output.

4.3.1 English processing in Chinese to English single system translation in news area

In the English translation output, we first removed the un-translated Chinese characters. Then, we applied a simple heuristic to normalize punctuation, capitalization and contractions in the translation output.

4.3.2 Chinese processing in the two English to Chinese machine translation tasks

The un-translated English words in the Chinese translation output in the two English to Chinese machine translation tasks were kept in the output. Simple heuristic was applied to remove the leading spaces in the Mongolian translation output.

4.3.3 Mongolian processing in Chinese to Mongolian machine translation on daily expressions task

In the Mongolian translation output, the un-translated Chinese characters were kept in the output. Simple heuristic was applied to remove the leading spaces in the Mongolian translation output.

4.4 Results

The official results were automatically evaluated using BLEU-SBP (Chiang et al., 2008), BLEU (Papineni et al., 2002), NIST (Doddington, 2002), GTM (Turian, 2003), WER, PER (Tillmann et al., 1997) and ICT.

Table 3: Official evaluation results on the Chinese to English single system translation in news area

Evaluation Task	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT
Chinese-English news	0.1488	0.1592	6.2476	0.6527	0.8340	0.5515	0.2825

4.4.1 Chinese to English single system translation in news area

Table 3 shows the official evaluation result on the Chinese to English single system translation in news area. We achieved a BLEU-SBP score of 0.1488.

Table 4: Official evaluation results on the English to Chinese machine translation in news area

Evaluation Task	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER
English-Chinese news	0.2410	0.2543	0.1925	8.5770	8.5815	0.7454	0.7936	0.4295

4.4.2 English to Chinese machine translation in news area

Table 4 shows the official evaluation result on the English to Chinese machine translation in news area. We achieved a BLEU-SBP score of 0.2410.

Table 5: Official evaluation results on the English to Chinese machine translation in sci-tech area

Evaluation Task	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT
English-Chinese sci-tech	0.3883	0.3924	0.3237	10.3937	10.4077	0.8774	0.6340	0.2938	0.4874

4.4.3 English to Chinese machine translation in science and technology area

Table 5 shows the official evaluation result on the English to Chinese machien translation in science and technology area. We achieved a BLEU-SBP score of 0.3883.

Table 6: Official evaluation results on the Chinese to Mongolian machine translation on daily expressions

Evaluation Task	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT
Chinese-Mongolian daily expression	0.1445	0.1564	4.9564	0.5717	0.6591	0.5593	0.4594

4.4.4 Chinese to Mongolian machine translation on daily expressions

Table 6 shows the official evaluation result on the Chinese to Mongolian machine translation on daily expressions. We achieved a BLEU-SBP score of 0.1445.

5 Conclusions

We have described experiments conducted at HKUST during the CWMT 2009 evaluation campaign on machine translation. We have reported the results we achieved on the four tasks we participated.

References

David Chiang, Steve DeNeefe, Yee Seng Chan, and Hwee Tou Ng. Decomposability of translation metrics for improved evaluation and efficient algorithms. In Proceedings of Empirical Methods in Natural Language Processing (EMNLP2008). Honolulu, Hawaii, 2008.

George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Human Language Trchnology conference (HLT-2002).* San Diego, CA, 2002.

Pascale Fung, Grace Ngai, Yongsheng Yang, and Benfeng Chen. A Maximum-Entorpy Chinese Parser augmented by transformation-based learning. ACM Transactions on Asian Language Information Processing (TALIP), 3(2). 2004. Pages 159-168.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL) Companion Volume Proceedings of the Demo and Poster Sessions*. Prague, Czech Republic, June 2007. Pages 177-180.

Philipp Koehn. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *6th Conference of the Association for Machine Translation in the Americas (AMTA)*. Washington, DC. September 2004.

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

Yihai Shen, Chi-kiu Lo, Marine Carpuat, Dekai Wu. HKUST Statistical Machine Translation Experiments for IWSLT 2007. In Proceedings of International Workshop on Spoken Language Translation (IWSLT). 2007. Pages 84-88.

Andreas Stolcke. SRILM – An extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*. Denver, Colorado. September 2002.

Christoph Tillmann, Stephan Vogel, Hermann Ney, A. Zubiaga, and H. Sawaf. Accerlerated DP-based search for statistical translation. In Eurospeech'97. Rhodes, Greece, 2997. Pages 2667-2670.

Joseph P. Turian, Luke Shen, and I. Dan Melamed. Evaluation of Machine Translation and its Evaluation, Proteus technical report #03-005. MT Summit IX, New Orleans, LA. 2003.

Ming Zhou, Bo Wang, Shujie Liu, Mu Li, Dongdong Zhang and Tiejun Zhao. Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points. In *Proceedings of Computational Linguistics (Coling)*. 2008. Pages 1121-1128.

Andreas Zollmann, Ashish Venugopal, Stephan Vogel, and Alex Waibel. The CMU-UKA syntax augmented machine translation system for IWSLT06. In Proceedings of the International Workshop on Spoken Language Translation. Kyoto, Japan. 2006. Pages 138-144.