HKUST Statistical Machine Translation Experiments for IWSLT 2007

Yihai SHEN Chi-kiu LO Marine CARPUAT Dekai WU

Human Language Technology Center
HKUST
Department of Computer Science, University of Science and Technology, Hong Kong
{shenyh,jackielo,marine,dekai}@cs.ust.hk

Abstract

This paper describes the HKUST experiments in the IWSLT 2007 evaluation campaign on spoken language translation. Our primary objective was to compare the open-source phrase-based statistical machine translation toolkit Moses against Pharaoh. We focused on Chinese to English translation, but we also report results on the Arabic to English, Italian to English, and Japanese to English tasks.

1. Introduction

We describe experiments conducted at HKUST during the IWSLT 2007 evaluation campaign on spoken language translation. For our second participation in the IWSLT evaluation, our focus was on experimenting with Moses (Koehn *et al.*, 2007), the new open-source toolkit for phrase-based Statistical Machine Translation (SMT), and on comparing it against its closed-source predecessor Pharaoh (Koehn, 2004) which we used in our IWSLT 2006 submission (Carpuat *et al.*, 2006).

Our main focus was on the Chinese-English task, which, this year, used clean text as opposed to the other tasks where speech transcriptions were to be translated. We also report results on all the language pairs, although we did not do any tuning or any language-specific processing for the Arabic to English, Japanese to English and Italian to English tasks.

*This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under GALE Contract No. HR0011-06-C-0023, and by the Hong Kong Research Grants Council (RGC) research grants RGC6083/99E, RGC6256/00E, and DAG03/04.EG09. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Defense Advanced Research Projects Agency.

2. Machine translation engine

We focus on phrase-based statistical machine translation since this approach has been shown to achieve competitive translation quality in many state-of-the-art systems (e.g., Zens *et al.* (2005), Shen *et al.* (2006), Eck *et al.* (2006)), and therefore is a good benchmark to evaluate decoder architectures.

2.1. Decoder

We use the Moses decoder (Koehn *et al.*, 2007), which is an open source toolkit for statistical machine translation. Just like the Pharaoh decoder, Moses uses a log-linear model, which combines several knowledge sources in translation decisions. Moses differs from Pharaoh by its represention of each input word as a factor as opposed to the word suface form only. Factors can include additional information such as part-of-speech, class, morphology and allow the phrase-based model to incorporate richer linguistic information. However, we do not use the factored representation in this first set of experiments, and use the surface form of words, just like in Pharaoh.

2.2. Phrasal bilexicon

The core phrasal bilexicon is obtained by collecting phrase pairs that are consistent with the IBM model 4 alignments obtained with GIZA++ (Och and Ney, 2002).

During phrase extraction, we tried two different methods to get the final word alignment from the bi-directional GIZA++ alignments: (1) intersect and (2) grow-diag-final. Intersect uses the strict intersection of the bidirectional word alignments, while grow-diag-final expands the alignment by adding directly neighboring alignment points, and alignment points in the diagonal neighborhood. We found that using grow-diag-final improves

Table 1: Resegmenting test sentences improves BLEU score.

IWSLT-07 data set	# original sentences	# sentences after resegmentation	BLEU with original sentences	BLEU after resegmentation
CE devtest1	506	546	41.09	**42.05**
CE devtest2	500	543	42.43	**43.76**
CE devtest3	506	558	51.68	**53.51**

Table 2: IWSLT-07 Training data statistics computed for the 4 language pairs

Training data statistics	Chinese-English	Arabic-English	Italian-English	Japanese-English
Number of bisentences	39953	19972	19972	39953
Vocabulary size (input lang)	11178	25152	17917	12535
Vocabulary size (English)	18992	13337	13337	18992

BLEU by around 0.5 over using intersect on the development test sets.

Since the IWSLT data consists of short sentences in a well-defined domain, we learn relatively long phrasal translations, allowing phrases of length up to 20 words.

2.3. Language model

The language models used in the translation tasks are trained with Kneser-Ney smoothing using the SRI language modeling toolkit (Stolcke, 2002). For the Arabic, Italian and Japanese tasks, the default language model was a trigram. While for the Chinese to English task, we used a 4-gram model and also an additional trigram model trained on the English gigaword corpus.

3. Chinese-English experimental setup

3.1. Data description

The training set is exactly the same as for the IWSLT 2006 evaluation campaign. They contain relatively short sentences from simple conversations in the travel domain drawn from the multilingual *Basic Travel Expression Corpus (BTEC)*.

The development test sets are the same as for IWSLT 2006, with the addition of the evaluation test data from IWSLT 2006. The first three development test set, the training data are clean text and the remaining two are transcriptions of the output of speech recognition systems.

This year, the Chinese-English test data consisted of clean text only. This contrast from previous evaluations where manual and automatic transcriptions of speech had to be translated. The IWSLT 2007 test set therefore matches more closely with the training data and the first three development test data, as opposed to the more recent Chinese-English tasks where automatic transcriptions of read and spontaneous speech were used.

3.2. Training data preprocessing

For the training data, we used the same basic preprocessing as in our IWSLT 2006 submission, which consists in performing tokenization and case normalization. The case normalization method is the same as the one described in Zollmann *et al.* (2006), where the first word of the sentence is normalized to its most frequent form.

English: The English was simply tokenized and case-normalized in the same manner for all languages.

Chinese: We use the LDC segmenter to re-segment the Chinese side of the corpus to get a better segmentation.

3.3. English text normalization

For all language pairs, in addition to training data normalization, we use simple heuristics to normalize punctuation, capitalization and contractions in the English output.

3.4. Improved sentence segmentation

Since the training data is drawn from clean text as opposed to speech transcriptions, we took advantage of the punctuation to resegment parallel sentences. This improves the quality of the word alignment, of the bilexicon and of the final translations.

Since the Chinese-English evaluation test set also used clean text, simple heuristics could also be used at testing time to improve the sentence segmentation of the decoder input. This prohibits incorrect long dis-

Table 3: Official BLEU scores for all translation tasks

IWSLT-07 task	Clean Transcription	ASR Output
Chinese-English	34.26	N/A
Arabic-English	19.51	14.20
Italian-English	17.02	17.02
Japanese-English	40.51	32.49

Table 4: Evaluation of translation quality on the Chinese to English translation task (clean transcriptions only)

IWSLT-07 data set	BLEU	NIST	METEOR	METEOR no synonyms	TER	WER	PER	CDER
CE devtest1 (buggy)	45.49	7,78	66.11	64.50	36.13	41.68	36.25	37.10
CE devtest1	46.32	8.00	68.01	66.41	36.18	41.35	36.12	37.14
CE devtest2 (buggy)	48.23	8.32	68.98	67.22	34.99	40.78	34.45	35.43
CE devtest2	49.77	8.82	71.88	69.85	34.47	40.12	33.41	34.58
CE devtest3 (buggy)	56.44	9.26	76.57	74.47	29.40	34.16	28.86	33.02
CE devtest3	58.29	9.61	78.48	76.28	28.29	32.67	27.62	29.15
CE test (buggy)	34.04	6.18	58.28	56.50	45.53	49.15	44.17	41.53
CE test	35.12	6.51	60.47	58.57	44.89	48.30	43.40	41.50

tance reorderings, and therefore significantly improves the translation quality of long sentences. Table 1 reports the impact of test sentence resegmentation on BLEU score under a preliminary experimental setting.

4. Experimental setup for other language pairs

For all other language pairs, we simply ran the system without any tuning, and with only minimal language specific processing.

Training corpus statistics for all the additional tasks can be found in Table 2. The Japanese-English training corpora contain twice as many sentences as the Arabic-English and Italian-English corpora. The English side of Arabic-English and Italian-English parallel text is a subset of that available for Chinese-English and Japanese-English.

We only performed basic tokenization for Arabic, Italian and Japanese. No sentence resegmentation nor additional normalization were performed for these tasks.

Arabic: In contrast with the 4 other languages considered, Arabic is a morphologically rich language and requires more sophisticated processing. The Arabic text is first converted to the Buckwalter romanization scheme. Tokenization and lemmatization are performed using the ASVMT Arabic morphological analysis toolkit (Diab, 2005). An Arabic word is typically formed of a stem, and possibly affixes and clitics. Affixes are inflectional markers for tense, gender and/or number, while the clitics include some prepositions, conjunctions, determiners, etc. Tokenization, which consists of separating those syntactic units, is the first step of processing in ASVMT. This is followed by lemmatization which, in ASVMT, refers to a normalization step where the tokens coming from stems that were modified when agglutinated are converted back to their original form.

Italian: We preprocessed the Italian corpus just like the English corpus: it was simply tokenized, using the same rules as for English, and case-normalized. This is obviously not optimal, as Italian presents more morphological inflexions than English, as suggested by the larger vocabulary size on the Italian side of the training data than on the English side (Table 2.)

Japanese: We used the provided word segmentation and did not perform any additional processing.

5. Experimental results

The official BLEU scores for HKUST's submitted runs, which were buggy due to accidental errors in combining the models and parameters used in the experiments, are shown in Table 3 for all four language pairs. The official results were only automatically evaluated using BLEU score (Papineni *et al.*, 2002). We achieved a BLEU score of 34.26 on Chinese to English read speech translation. There were 9 primary submissions to that task, with BLEU scores ranging from 19.34 to 40.77.

Our subsequent debugged runs yielded higher trans-

lation accuracy. Updated results for our debugged runs on the development sets are reported in Table 4 for the Chinese-English task. For running the submitted buggy model on the official IWSLT-07 test set, there is a slight difference between the official BLEU score of 34.26 and our own measurement of 34.04. This difference appears to be caused by slight differences between BLEU scoring tools and settings (the tool we are using appears to give lower scores).We also computed the other most commonly used automatic evaluation metrics for translation quality: NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005) and Translation Edit Rate (TER) (Snover *et al.*, 2006), Word Error Rate (WER), Position-Independent Word Error Rate (PER) and CDER (Leusch *et al.*, 2006).

6. Comparing Moses results with Pharaoh

Using the same phrasal bilexicon and language model as with Moses, we performed several contrastive runs using Pharaoh, all other settings being identical. Results are reported in Table 5 for three different baseline experimental settings. We performed many experimental runs in which we vary the experimental settings and pre or post processing steps, e.g. phrase tables, language models, to compare the translation quality produced by Pharoah and Moses. The BLEU scores show that Moses almost always outperforms Pharaoh.

7. Contrastive experiments: WSD and ITG decoding

We also contrasted the Moses results with other decoder architectures that are quite different from both Moses itself and Pharaoh, although this was not our focus in this set of experiments, and work on each decoder was done independently.

The first contrastive decoder is the Word Sense Disambiguation (WSD) augmented system described in Carpuat and Wu (2007). The phrasal bilexicon is augmented with dynamically computed context-sensitive translation probabilities, based on traditional WSD models trained on the phrase-aligned parallel text. In this paper, we used WSD to augment Moses instead of Pharaoh.

The second contrastive decoder is based on a Bracketing Inversion Transduction Grammar (BITG) model, as described in Wu (1996) and Wu (1997).

The BLEU scores show that in the runs where the WSD-augmented Moses model or the BITG decoder are available, they outperform both Moses and Pharaoh.

Table 5: BLEU scores for a representative sample of decoding runs using Moses vs. Pharaoh and other decoder

Run No.	**Pharaoh**	**Moses**	**WSD**	**BITG**
1	41.14	**41.17**		
2	41.65	41.70	**43.47**	
3	42.05	42.16		**43.04**
4	43.40	**43.55**		
5	41.92	**42.26**		
6	42.80	**43.19**		
7	43.76	**44.28**		
8	44.17	**44.64**		
9	51.64	**52.19**		
10	52.15	**52.59**		
11	53.51	**53.64**		
12	**53.87**	53.53		

8. Conclusion

We have described experiments conducted at HKUST during the IWSLT 2007 evaluation campaign on spoken language translation, focussing primarily on the Chinese to English translation task. The open-source Moses-based decoder system achieves slightly higher performance on average than its closed-source predecessor Pharaoh, and therefore constitutes a reasonable baseline for further experiments with other decoder architectures, such as the promising contrastive experiments with HKUST architectures utilizing WSD-augmented and ITG-based decoding.

References

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgement. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005)*, Ann Arbor, Michigan, June 2005.

Marine Carpuat and Dekai Wu. Improving statistical machine translation using word sense disambiguation. In *The 2007 Joint Conference on Empirical Methods in Natural Language Processing and Com-*

putational Natural Language Learning (EMNLP-CoNLL 2007), Prague, June 2007.

Marine Carpuat, Yihai Shen, Xiaofeng Yu, and Dekai Wu. Toward integrating word sense and entity disambiguation into statistical machine translation. In *Third International Workshop on Spoken Language Translation (IWSLT 2006)*, Kyoto, November 2006.

Mona Diab. Documentation for the Arabic SVM Toolkit. http://www.cs.columbia.edu/mdiab/, 2005.

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

Matthias Eck, Ian Lane, Nguyen Bach, Sanjika Hewavitharana, Muntsin Kolss, Bing Zhao, Almut Silja Hildebrand, Stephan Vogel, and Alex Waibel. The UKA/CMU Statistical Machine Translation System for IWSLT 2006. In *Proc. of the International Workshop on Spoken Language Translation*, pages 130–137, Kyoto, Japan, 2006.

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 Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

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.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. CDER: Efficient MT evaluation using block movements. In *11st Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, Trento, Italy, April 2006.

Franz Josef Och and Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th meeting of the Association for Computational Linguistics (ACL-2002)*, Philadelphia, 2002.

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 for Computational Linguistics*, 2002.

Wade Shen, Richard Zens, Nicola Bertoldi, and Marcello Federico. The JHU Workshop 2006 IWSLT System. In *Proc. of the International Workshop on Spoken Language Translation*, pages 59–63, Kyoto, Japan, 2006.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *AMTA*, pages 223–231, Boston, MA, 2006. Association for Machine Translation in the Americas.

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

Dekai Wu. A polynomial-time algorithm for statistical machine translation. In *Proceedings of 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, California, June 1996.

Dekai Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–404, 1997.

Richard Zens, Oliver Bender, Sasa Hasan, Shahram Khadivi, Evgeny Matusov, Jia Xu, Yuqi Zhang, and Hermann Ney. The RWTH phrase-based statistical machine translation system. In *Proceedings of IWSLT-2005*, Pittsburgh, PA, 2005.

Andreas Zollmann, Ashish Venugopal, Stephan Vogel, and Alex Waibel. The CMU-UKA Syntax Augmented Machine Translation System for IWSLT-06. In *Proc. of the International Workshop on Spoken Language Translation*, pages 138–144, Kyoto, Japan, 2006.