Recent Improvements in the CMU Large Scale Chinese-English SMT System

Almut Silja Hildebrand, Kay Rottmann, Mohamed Noamany, Qin Gao, Sanjika Hewavitharana, Nguyen Bach and **Stephan Vogel**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
silja, kayrm, mfn, qing, sanjika, nbach, vogel+@cs.cmu.edu

Abstract

In this paper we describe recent improvements to components and methods used in our statistical machine translation system for Chinese-English used in the January 2008 GALE evaluation. Main improvements are results of consistent data processing, larger statistical models and a POS-based word reordering approach.

1 Introduction

Building a full scale Statistical Machine Translation (SMT) system involves many preparation and training steps and it consists of several components, each of which contribute to the overall system performance. Between 2007 and 2008 our system improved by 5 points in BLEU from 26.60 to 31.85 for the unseen MT06 test set, which can be mainly attributed to two major points.

The fast growth of computing resources over the years make it possible to use larger and larger amounts of data in training. In Section 3 we show how parallelizing model training can reduce training time by an order of magnitude and how using larger training data as well as more extensive models improve translation quality.

Word reordering is still a difficult problem in SMT. In Section 4 we apply a Part Of Speech (POS) based syntactic reordering model successfully to our large Chinese system.

1.1 Decoder

Our translation system is based on the CMU SMT decoder as described in (Hewavitharana et al., 2005). Our decoder is a phrase-based beam search decoder, which combines multiple models e.g. phrase tables, several language models, a distortion model ect. in a log-linear fashion. In order to find an optimal set of weights, we use MER training as described in (Venugopal et al., 2005), which uses rescoring of the top n hypotheses to maximize an evaluation metric like BLEU or TER.

1.2 Evaluation

In this paper we report results using the BLEU metric (Papineni et al., 2002), however as the evaluation criterion in GALE is HTER (Snover et al., 2006), we also report in TER (Snover et al., 2005).

We used the test sets from the NIST MT evaluations from the years 2003 and 2006 as development and unseen test data.

1.3 Training Data

In translation model training we used the Chinese-English bilingual corpora relevant to GALE available through the LDC[1]. After sentence alignment these sources add up to 10.7 million sentences with 301 million running words on the English side. Our preprocessing steps include tokenization on the English side and for Chinese: automatic word segmentation using the revised version of the Stanford Chinese Word Segmenter[2] (Tseng et al., 2005) from 2007, replacement of traditional by simplified Chinese characters and 2-byte to 1-byte ASCII character normalization. After data cleaning steps like e.g. removal of sentence pairs with very unbalanced sen-

[1] http://projects.ldc.upenn.edu/gale/data/catalog.html
[2] http://nlp.stanford.edu/software/segmenter.shtml

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 77–80,
Columbus, Ohio, USA, June 2008.

tence length etc., we used the remaining 10 million sentences with 260 million words (English) in translation model training (260M system).

2 Number Tagging

Systematic tagging and pre-translation of numbers had shown significant improvements for our Arabic-English system, so we investigated this for Chinese-English. The baseline for these experiments was a smaller system with 67 million words (67M) bilingual training data (English) and a 500 million word 3-gram LM with a BLEU score of 27.61 on MT06. First we pre-translated all numbers in the testdata only, thus forcing the decoder to treat the numbers as unknown words. Probably because the system could not match longer phrases across the pre-translated numbers, the overall translation quality degraded by 1.6 BLEU to 26.05 (see Table 1).

We then tagged all numbers in the training corpus, replaced them with a placeholder tag and re-trained the translation model. This reduced the vocabulary and enabled the decoder to generalize longer phrases across numbers. This strategy did not lead to the expected result, the BLEU score for MT06 only reached 25.97 BLEU.

System	MT03	MT06
67M baseline	31.45/60.93	27.61/62.18
test data tagged	–	26.06/63.36
training data tagged	29.07/62.52	25.97/63.39

Table 1: Number tagging experiments, BLEU/TER

Analysing this in more detail, we found, the reason for this degradation in translation quality could be the unbalanced occurrence of number tags in the training data. From the bilingual sentence pairs, which contain number tags, 66.52% do not contain the same number of tags on the Chinese and the English side. As a consequence 52% of the phrase pairs in the phrase table, which contain number tags had to be removed, because the tags were unbalanced. This hurts system performance considerably.

3 Scaling up to Large Data

3.1 Language Model

Due to the availability of more computing resources, we were able to extend the language model history from 4- to 5-gram, which improved translation quality from 29.49 BLEU to 30.22 BLEU for our large scale 260M system (see Table 2). This shows, that longer LM histories help if we are able to use enough data in model training.

System	MT03	MT06
260M, 4gram	31.20/61.00	29.49/61.00
260M, 5gram	32.20/60.59	30.22/60.81

Table 2: 4- and 5-gram LM,260M system, BLEU/TER

The language model was trained on the sources from the English Gigaword Corpus V3, which contains several newspapers for the years between 1994 to 2006. We also included the English side of the bilingual training data, resulting in a total of 2.7 billion running words after tokenization.

We trained separate open vocabulary language models for each source and interpolated them using the SRI Language Modeling Toolkit (Stolcke, 2002). Table 3 shows the interpolation weights for the different sources. Apart from the English part of the bilingual data, the newswire data from the Chinese Xinhua News Agency and the Agence France Press have the largest weights. This reflects the makeup of the test data, which comes in large parts from these sources. Other sources, as for example the UN parlamentary speeches or the New York Times, differ significantly in style and vocabulary from the test data and therefore get small weights.

xin 0.30	cna 0.06	nyt 0.03
bil 0.26	un 0.07	ltw 0.01
afp 0.21	apw 0.05	

Table 3: LM interpolation weights per source

3.2 Speeding up Model Training

To accelerate the training of word alignment models we implemented a distributed version of GIZA++ (Och and Ney, 2003), based on the latest version of GIZA++ and a parallel version developed at Peking University (Lin et al., 2006). We divide the bilingual training data in equal parts and distribute it over several processing nodes, which perform alignment independently. In each iteration the nodes read the model from the previous step and output all necessary counts from the data for the models, e.g. the

co-occurrence or fertility model. A master process collects the counts from the nodes, normalizes them and outputs the intermediate model for each iteration.

This distributed GIZA++ version finished training the word alignment up to IBM Model 4 for both language directions on the full bilingual corpus (260 million words, English) in 39 hours. On average about 11 CPUs were running concurrently. In comparison the standard GIZA++ implementation finished the same training in 169 hours running on 2 CPUs, one for each language direction.

We used the Pharaoh/Moses package (Koehn et al., 2007) to extract and score phrase pairs using the grow-diag-final extraction method.

3.3 Translation Model

We trained two systems, one on the full data and one without the out-of-domain corpora: UN parlament, HK hansard and HK law parallel texts. These parlamentary sessions and law texts are very different in genre and style from the MT test data, which consists mainly of newspaper texts and in recent years also of weblogs, broadcast news and broadcast conversation. The in-domain training data had 3.8 million sentences and 67 million words (English). The 67 million word system reached a BLEU score of 29.65 on the unseeen MT06 testset. Even though the full 260M system was trained on almost four times as many running words, the baseline score for MT06 only increased by 0.6 to 30.22 BLEU (see Table 4).

System	MT03	MT06
67M in-domain	32.42/60.26	29.65/61.22
260M full	32.20/60.59	30.22/60.81

Table 4: In-domain only or all training data, BLEU/TER

The 67M system could not translate 752 Chinese words out of 38937, the number of unknown words decreased to 564 for the 260M system. To increase the unigram coverage of the phrase table, we added the lexicon entries that were not in the phrase table as one-word translations. This lowered the number of unknown words further to 410, but did not effect the translation score.

4 POS-based Reordering

As Chinese and English have very different word order, reordering over a rather limited distance during decoding is not sufficient. Also using a simple distance based distortion probability leaves it essentially to the language model to select among different reorderings. An alternative is to apply automatically learned reordering rules to the test sentences before decoding (Crego and Marino, 2006). We create a word lattice, which encodes many reorderings and allows long distance reordering. This keeps the translation process in the decoder monotone and makes it significantly faster compared to allowing long distance reordering at decoding time.

4.1 Learning Reordering Rules

We tag both language sides of the bilingual corpus with POS information using the Stanford Parser[3] and extract POS based reordering patterns from word alignment information. We use the context in which a reordering pattern is seen in the training data as an additional feature. Context refers to the words or tags to the left or to the right of the sequence for which a reordering pattern is extracted.

Relative frequencies are computed for every rule that has been seen more than n times in the training corpus (we observed good results for $n > 5$).

For the Chinese system we used only 350k bilingual sentence pairs to extract rules with length of up to 15. We did not reorder the training corpus to retrain the translation model on modified Chinese word order.

4.2 Applying Reordering Rules

To avoid hard decisions, we build a lattice structure for each source sentence as input for our decoder, which contains reordering alternatives consistent with the previously extracted rules.

Longer reordering patterns are applied first. Thereby shorter patterns can match along new paths, creating short distance reordering on top of long distance reordering. Every outgoing edge of a node is scored with the relative frequency of the pattern used on the following sub path (For details see (Rottmann and Vogel, 2007)). These model scores give this re-

[3] http://nlp.stanford.edu/software/lex-parser.shtml

ordering approach an advantage over a simple jump model with a sliding window.

System	MT03	MT06
260M, standard	32.20/60.59	30.22/60.81
260M, lattice	33.53/59.74	31.74/59.59

Table 5: Reordering lattice decoding in BLEU/TER

The system with reordering lattice input outperforms the system with a reordering window of 4 words by 1.5 BLEU (see Table 5).

5 Summary

The recent improvements to our Chinese-English SMT system (see Fig. 1) can be mainly attributed to a POS based word reordering method and the possibility to work with larger statistical models.

We used the lattice translation functionality of our decoder to translate reordering lattices. They are built using reordering rules extracted from tagged and aligned parallel data. There is further potential for improvement in this approach, as we did not yet reorder the training corpus and retrain the translation model on modified Chinese word order.

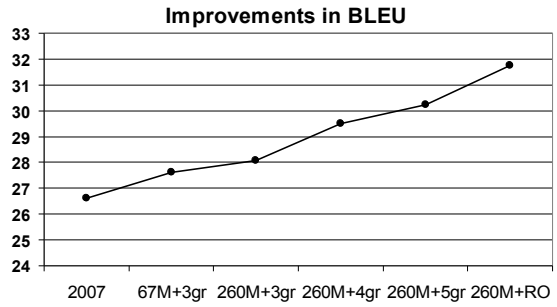

Figure 1: Improvements for MT06 in BLEU

We modified GIZA++ to run in parallel, which enabled us to include especially longer sentences into translation model training. We also extended our decoder to use 5-gram language models and were able to train an interpolated LM from all sources of the English GigaWord Corpus.

Acknowledgments

This work was partly funded by DARPA under the project GALE (Grant number #HR0011-06-2-0001).

References

Josep M. Crego and Jose B. Marino. 2006. Reordering Experiments for N-Gram-Based SMT. *Spoken Language Technology Workshop*, Palm Beach, Aruba.

Sanjika Hewavitharana, Bing Zhao, Almut Silja Hildebrand, Matthias Eck, Chiori Hori, Stephan Vogel and Alex Waibel. 2005. The CMU Statistical Machine Translation System for IWSLT 2005. *IWSLT 2005*, Pittsburgh, PA.

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, Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *ACL 2007, Demonstration Session*, Prague, Czech Republic.

Xiaojun Lin, Xinhao Wang, and Xihong Wu. 2006. NLMP System Description for the 2006 NIST MT Evaluation. *NIST 2006 MT Evaluation*.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Poukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. *ACL 2002*, Philadelphia, USA.

Kay Rottmann and Stephan Vogel. 2007. Word Reordering in Statistical Machine Translation with a POS-based Distortion Model. *TMI-2007: 11th International Conference on Theoretical and Methodological Issues in MT,* Skvde, Sweden.

Mathew Snover, Bonnie Dorr, Richard Schwartz, John Makhoul, Linnea Micciula and Ralph Weischedel. 2005. A Study of Translation Error Rate with Targeted Human Annotation. *LAMP-TR-126*, University of Maryland, College Park and BBN Technologies.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. *7th Conference of AMTA*, Cambridge, Massachusetts, USA.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. *ICSLP*, Denver, Colorado.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky and Christopher Manning. 2005. A Conditional Random Field Word Segmenter. *Fourth SIGHAN Workshop on Chinese Language Processing*.

Ashish Venugopal, Andreas Zollman and Alex Waibel. 2005. Training and Evaluation Error Minimization Rules for Statistical Machine Translation. *ACL 2005, WPT-05, Ann Arbor, MI*