The Johns Hopkins University 2003
Chinese-English Machine Translation System

W. Byrne, S. Khudanpur, W. Kim, S. Kumar, P. Pecina, P. Virga, P. Xu and D. Yarowsky

Center for Language and Speech Processing, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD, 21218, USA

Abstract
We describe a Chinese to English Machine Translation system developed at the Johns Hopkins University for the NIST 2003 MT evaluation. The system is based on a Weighted Finite State Transducer implementation of the alignment template translation model for statistical machine translation. The baseline MT system was trained using 100,000 sentence pairs selected from a static bitext training collection. Information retrieval techniques were then used to create specific training collections for each document to be translated. This document-specific training set included bitext and name entities that were then added to the baseline system by augmenting the library of alignment templates. We report translation performance of baseline and IR-based systems on two NIST MT evaluation test sets.

1 Alignment Template Translation Model

We first give an outline of the Alignment Template Translation Model (ATTM) (Och, 2002) for statistical machine translation. The overall model is based on a two-level alignment between the source and the target sentence: a phrase-level alignment between source and target phrases and a word-level alignment between words in these phrase pairs implemented via individual alignment templates. The ATTM has been reformulated (Kumar and Byrne, 2003) so that both bitext word alignment and translation can be implemented using standard weighted finite state transducer (WFST) operations available from an AT&T FSM toolkit (Mohri et al., 1997).

The ATTM architecture is presented in Figure 1. The components of the overall translation model are the source language model, the source segmentation model, the phrase permutation model, the template sequence model, the phrasal translation model and the target language model. Each of these conditional distributions is modeled independently and implemented as a weighted finite state acceptor or transducer (Kumar and Byrne, 2003). In the implementation here, the ATTM maps Chinese word sequences to a sequence of English word classes, which are then mapped to English sentences.

2 Training and Test Data Sources

2.1 Bitext Training Data

Our bitext training set consisted of parallel corpora taken from 7 sources. These sources were the Chinese Treebank English parallel corpus, FBIS parallel text, Hong Kong News Parallel Text, Hong Kong Hansards Parallel Text, Sinorama Parallel Text, the United Nations Parallel Text and Xinhua Parallel News Text. All the sources are available from the LDC (NIST, 2003), and summarized in Table 1.

2.2 Test sets

Our test corpora consisted of two sets (NIST, 2003). The first corpus is the NIST MT 2001 dry-run test set (Dev02) consisting of 25 documents and 206 sentences. The second corpus is the Zaobao-news portion of the NIST MT 2002 evaluation set (ZBN-Eval02) consisting of 30 documents and 332 sentences. Both test sets contained four reference translations per Chinese source sentence. The statistics from the test sets are summarized in Table 1.

3 The Baseline System

3.1 Bitext Training Data

In building our baseline system, the FBIS Chinese-English parallel text (NIST, 2003) was used as the bitext...
training data. Since the FBIS data is aligned at the doc-
ument level, we performed sentence alignment for each
document pair using an aligner developed during JHU
WS’01 (Section 2.2.4 of (JHU, 2001)). Several succes-
sive filtering steps were implemented to deal with vari-
sious issues related to the baseline system. First of all, the
quality of the aligner output was not uniformly good. We
treated each sentence pair as two bags of words and com-
puted the average precision and recall of Chinese-English
word pair co-occurrence in a sentence pair, based on the
LDC Chinese English Translation Lexicon (versions 2
and 3) (LDC, 2002). All sentence pairs were then ranked
according to this score (Filter1). Secondly, we used a En-
lish text normalization tool developed during WS’99to
normalize the English text. The Chinese text was then
segmented by the LDC segmenter (LDC, 2002). Finally,
to speed up the translation model training, we put a length
constraint to discard all sentence pairs in which either
sentence is longer than 100 words. After all the steps,
we selected 100,000 sentence pairs as our final training
corpus. The first row of Table 2 summarizes the statistics
of the 100,000 sentence-pairs from FBIS data.

3.2 Bitext Word Alignments on training data

The alignment templates are based on bitext word
alignments on the training data. We obtained word align-
ments of bitext using IBM-4 translation models trained
in each translation direction (E→C and C→E) , and then
formed the union of these alignments (Och, 2002).

For IBM-4 model training, we augmented bitext
with word-pairs from the LDC Chinese-English dictio-



1http://www.english.people.com.cn

nary (LDC, 2002). A dictionary entry was added only if
both the English and the Chinese words occur in the bi-
text. Using this criterion, we selected 41,695 dictionary
entries and duplicated each entry 10 times (Och and Ney,
2000). IBM-4 translation models were then trained on
the resulting training text using the GIZA++ statistical
MT toolkit (Och, 2002).

3.3 Building the Alignment Template Library

We constructed the library of alignment templates
from the bitext word alignments using the phrase-extract
algorithm reported in Och (2002). This procedure iden-
tifies several alignment templates that are consistent with
a Chinese source phrase. To restrict the memory re-
quirements of the model, we extracted only the templates
which have at most 5 words in the source phrase. Fur-
thermore, we restricted ourselves to the templates which
have a relative frequency greater than 0.01.

We augmented the basic set of templates with three
additional types of templates. The first addition con-
sisted of phrasal entries (a Chinese-word mapping to a
English phrase) from the LDC dictionary (LDC, 2002)
The dictionary entries (10,183 entries) were included in
our template library. The second addition was a special-
ized rule based Chinese-to-English translator for num-
ders, dates and times. We first tagged numbers in the seg-
mented Chinese text and then translate the numbers af-
after normalizing them to a universal representation. These
translations were also included in our template library.
The third addition included templates that allow for in-
sertions of selected target words. All the target words
were ranked based on their probabilities of zero-fertility
in the IBM-4 word fertility model. We then selected the
top 20 words from this ranked list. This word list con-
sisted of 20 words that are primarily determiners, such
as “a”,“of” and “the”. Following this procedure, we ob-
tained templates based on Chinese words and English
words. We then modified the templates to allow all the
inflected forms of the English words.

3.4 Baseline Language Model for English

We trained a trigram word model from English news
text derived from two sources: online archives (Sept 1998
to Feb 2002) of The People’s Daily1 (16.9M words) and
the English side of the Xinhua Chinese-English parallel
Table 1: Statistics for the training and test sources.

| Source Corpus       | Doc Pairs | Sentence Pairs | Unique Sentences | Words English | Vocabulary English | Vocabulary Chinese |
|---------------------|-----------|----------------|------------------|--------------|--------------------|-------------------|
| Training            | 325       | 3,464          | 3,190            | 3,208        | 100,361            | 139,379           |
| Ch Treebank         | 11,537    | 253,555        | 232,178          | 237,207      | 8,449,546          | 11,006,282        |
| FBIS                | 194       | 380,437        | 348,165          | 352,409      | 11,487,018         | 13,752,213        |
| HKHansards          | 18,147    | 218,099        | 190,440          | 191,952      | 6,796,094          | 7,392,625         |
| HKNews              | 2,373     | 107,141        | 106,458          | 106,949      | 3,395,656          | 3,928,678         |
| Sinorama            | 44,754    | 3,210,712      | 3,022,758        | 2,997,876    | 105,124,525        | 121,881,108       |
| UN                  | 19,140    | 121,881        | 118,363          | 119,705      | 4,111,915          | 4,258,744         |
| Sinorama            | 44,754    | 3,210,712      | 3,022,758        | 2,997,876    | 105,124,525        | 121,881,108       |
| Xinhua              | 19,140    | 121,881        | 118,363          | 119,705      | 4,111,915          | 4,258,744         |
| Total               | 96,470    | 4,295,289      | 4,012,434        | 3,998,255    | 139,465,115        | 162,359,029       |

Test

| Source Corpus       | Doc Pairs | Sentence Pairs | Unique Sentences | Words English | Vocabulary English | Vocabulary Chinese |
|---------------------|-----------|----------------|------------------|--------------|--------------------|-------------------|
| Dev02               | 25        | 206            | 206              | 5,582        | 1,683              |                   |
| ZBN-Eval02          | 30        | 332            | 332              | 8,533        | 2,621              |                   |

Table 2: Final Training data statistics for the Baseline and the IR systems. *Statistics for the document-specific training sets were averaged over all the test documents.

| Source Corpus       | Dev02 | ZBN-Eval02 |
|---------------------|-------|------------|
| Chinese Treebank    | 0     | 0.02       |
| FBIS                | 9.84  | 2.89       |
| HKHansards          | 36.36 | 47.02      |
| HKNews              | 2.96  | 0.24       |
| Sinorama            | 2.05  | 0.35       |
| UN                  | 48.39 | 49.45      |
| Xinhua              | 0.39  | 0.03       |

Table 3: Contribution (%) of sources of sentence-pairs averaged over the documents in each test set.

4 The IR based system

We now describe a second translation system that was trained on bitext data selected from the seven bitext sources using information retrieval techniques.

4.1 Document Specific training bitexts

For each test document we created a specific bitext training set. We employed a standard Information Retrieval vector model (Baeza-Yates and Ribeiro-Neto, 1999). Chinese documents from the test set and from all training text sources were represented as vectors, and the cosine distance between those vectors represented the degree of similarity between each test document and every training set document. Index terms were both Chinese words and characters (Nie and Ren, 1999); stopwords were not used, and term weights were calculated simply as raw relative frequencies of words in the document.

For each test document the training set was filtered based on similarity scores, sentence-alignment score (Section 3.1) (≤ 0.35) and length (> 60 words). The final training text for each document to be translated contained approximately 100,000 sentence-pairs from the documents with high similarity scores (Tables 2 and 3).

4.2 Document Specific Translation Models

In these experiments, we first trained IBM-4 translation models in both translation directions on the training subsets that have been found to be relevant to each test document. We merged the word alignments on the baseline FBIS bitext with the alignments found from the document specific bitext collection, and then extracted alignment templates specialized for each test document. This generated N different template libraries and vocabularies for the N test documents.

4.3 Incorporation of Name Entities (NEs)

We used the LDC Chinese-English Name Entity Lists (NIST, 2003) to identify NEs in the test documents. Rather than including the entries from the NE lists in the
segmenter lexicon and performing a new word segmentation of Chinese, we took an alternate approach made possible by the ATTM. In our approach, we used all of the data sources (Chinese text segmented) as the “universe”. For each test document, we first retrieved Chinese documents from the universe that had a cosine similarity score greater than 0.65; these were identified as documents that potentially have the same NEs as in the test document. All the English names that appeared in the corresponding English documents were identified using the LDC NE lists, together with all of their possible Chinese translations. We then filtered the resulting list by discarding any entry whose Chinese part (as a Chinese character sequence) was not in the retrieved Chinese documents. For those that did appear, we preserved the segmentations from the retrieved documents. This approach allowed us to pick NEs which were not initially segmented as a single word, and to make an NE list that maps a Chinese “phrase” to a single English word. The NE list was finally added to the ATTM as alignment templates (total of 11768 entries). A block diagram of the baseline and the IR systems is shown in Figure 2.

5 Translation Performance

We now present the translation performance of the baseline and the IR systems on the two development test sets. The translation performance was measured using the BLEU (Papineni et al., 2001) and the NIST MT-eval metrics (Doddington, 2002) using the four reference translation provided for each test sentence. The NIST and BLEU scores were measured using version 9 of the mteval software (NIST, 2003). We note that scaling factors such as Word Insertion Penalty and Grammar Scale factors were chosen appropriately for each test set. Also, the phrase segmentation model was also tuned to each test set. The pruned version of the language model was used to generate translation lattices which were then rescored with full language model to generate the final translation.

6 Conclusion

We have successfully demonstrated that Information Retrieval techniques can be used to construct training sets for statistical machine translation. Our initial experiments show gains over the baseline system. The IR approach allows us to identify relevant sentence translations as well as translation of name entities. The ATTM training and decoding framework allows a convenient way to incorporate these into the baseline system. Future work will involve refinements to the IR approach and better integration of the constituents into the ATTM framework.

Acknowledgments

We would like to thank F. J. Och of ISI/USC for useful discussions on the ATTM. We would also like to thank C. Schaefer of JHU for providing us the program to compute sentence-alignment scores and E. Drabek of JHU and R. Hwa of UMD for the program to translate Chinese numbers. We thank AT&T Labs - Research for use of the FSM Toolkit.

7 References

Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto, 1999. Modern Information Retrieval. ACM Press / Addison-Wesley.

G. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In Proc. of HLT 2002, San Diego, CA, USA.

2001. Automatic Summarization of Multiple (Multilingual) Documents. http://www.cslp.jhu.edu/ww2001/groups/asmd.

S. Kumar and W. Byrne. 2003. A weighted finite state transducer implementation of the alignment template model for statistical machine translation. In Proc. of the Conference on Human Language Technology, Edmonton, Canada.

LDC, 2002. LDC Chinese-English Dictionary and Chinese Segmenter. http://www.ldc.upenn.edu/Projects/Chinese.

M. Mohri, F. Pereira, and M. Riley, 1997. ATT General-purpose finite-state machine software tools. http://www.research.att.com/sw/tools/fsm/.

J.Y. Nie and F. Ren. 1999. Chinese information retrieval: using characters or words? Information Processing and Management, 35:443–462.

NIST. 2003. The nist machine translation evaluations. http://www.nist.gov/speech/tests/mt/.

F. Och and H. Ney. 2000. A comparison of alignment models for statistical machine translation. In Proc. Of 18th Conference On Computational Linguistics, pages 1086–1090, Saarbrucken, Germany.

F. Och. 2002. Statistical Machine Translation: From Single Word Models to Alignment Templates. Ph.D. thesis, RWTH Aachen, Germany.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division.

A. Stolcke. 2002. SRILM – an extensible language modeling toolkit. In Proc. of the International Conference on Spoken Language Processing, pages 901–904, Denver, CO, USA. http://www.speech.sri.com/projects/srilm/.

Table 4: Translation Performance. *The NE dictionary was not added to IR system on Dev02.