The XMU Phrase-Based Statistical Machine Translation System for IWSLT 2006

Yidong Chen, Xiaodong Shi
Institute of Artificial Intelligence
Xiamen University
P. R. China
November 28, 2006 - Kyoto
Outline

- Overview
- Training
- System
  - Translation Model
  - Parameters
  - Decoder
  - Dealing with the Unknown Words
  - Recovering the Missing Punctuations
  - Translating the ASR Lattice
- Experiments
- Conclusions
Overview

- **Who we are?**
  - NLP group at Institute of Artificial Intelligence, Xiamen University
  - Begin research on SMT since 2004
  - Have worked on rule-based MT for more than 15 years
  - First web MT in China (1999)
  - First mobile phone MT in China (2006)
  - Website: [http://ai.xmu.edu.cn/](http://ai.xmu.edu.cn/)
    [http://mt.xmu.edu.cn](http://mt.xmu.edu.cn)
    [http://nlp.xmu.edu.cn](http://nlp.xmu.edu.cn)
Overview (Cont.)

- **IWSLT 2006**
  - First participation
  - We implemented a simple **phrase-based** statistical machine translation system.
  - We participated in the **open data track** for **ASR lattice** and **Cleaned Transcripts** for the Chinese-English translation direction.
Outline

- Overview
- **Training**
- System
  - Translation Model
  - Parameters
  - Decoder
  - Dealing with the Unknown Words
  - Recovering the Missing Punctuations
  - Translating the ASR Lattice
- Experiments
- Conclusions
Training

- Preprocessing (Chinese part)
  - Segmentation
  - Mixed (DBC/SBC) case to SBC case
- Preprocessing (English part)
  - Tokenization
  - Truecasing of the first word of an English sentence
Training (Cont.)

- **Word Alignment**
  - Firstly, we ran **GIZA++** up to IBM model 4 in **both translation directions** to get an initial word alignment.
  - Then, We applied “**grow-diag-final**” method (Koehn, 2003) to refine it and achieve n-to-n word alignment.
Phrase Extraction

similar to (Och, 2002).

We limited the length of phrases from 1 word to 6 words.

For a Chinese phrase, only 20-best corresponding bilingual phrases were kept.

\[ \sum_{i=1}^{N} \lambda_i \cdot h_i(\tilde{e}, \tilde{c}) \] is used to evaluate and rank the bilingual phrases with the same Chinese phrase.
Training (Cont.)

- **Phrase Probabilities**
  - Phrase translation probability \( p(\tilde{e} | \tilde{c}) \)
  - Inversed phrase translation probability \( p(\tilde{c} | \tilde{e}) \)
  - Phrase lexical weight \( \text{lex}(\tilde{e} | \tilde{c}) \)
  - Inversed phrase lexical weight \( \text{lex}(\tilde{c} | \tilde{e}) \)

\[
p(\tilde{e} | \tilde{c}) = \frac{N(\tilde{e}, \tilde{c})}{\sum_{\tilde{e}'} N(\tilde{e}', \tilde{c})}
\]

\[
\text{lex}(\tilde{e} | \tilde{c}) = \text{lex}(e_i^I | c_j^J, a) = \prod_{i=1}^{I} \frac{1}{\{j | (i, j) \in a\}} \sum_{\forall(i, j) \in a} p(c_i | e_j)
\]
Outline

- Overview
- Training
- System
  - Translation Model
  - Parameters
  - Decoder
  - Dealing with the Unknown Words
  - Recovering the Missing Punctuations
  - Translating the ASR Lattice
- Experiments
- Conclusions
Translation Model

- We use a log-linear modeling (Och, 2002):

\[
\Pr(e^I_1 \mid c^J_1) = \frac{\exp[\sum_{m=1}^{M} \lambda_m \cdot h_m(e^I_1, c^J_1)]}{\sum_{e'_1} \exp[\sum_{m=1}^{M} \lambda_m \cdot h_m(e'_1, c^J_1)]}
\]

\[
\hat{e}^I_1 = \arg\max_{e'_1} \left\{ \sum_{m=1}^{M} \lambda_m \cdot h_m(e^I_1, c^J_1) \right\}
\]
Translation Model (Cont.)

- Seven features
  - Phrase translation probability \( p(\tilde{e} | \tilde{c}) \)
  - Inversed phrase translation probability \( p(\tilde{c} | \tilde{e}) \)
  - Phrase lexical weight \( \text{lex}(\tilde{e} | \tilde{c}) \)
  - Inversed phrase lexical weight \( \text{lex}(\tilde{c} | \tilde{e}) \)
  - English language model \( \text{lm}(e^I_1) \)
  - English sentence length penalty \( I \)
  - Chinese phrase count penalty \( -J' \)

- We didn’t use features on reordering.
Outline

- Overview
- Training
- System
  - Translation Model
  - Parameters
  - Decoder
  - Dealing with the Unknown Words
  - Recovering the Missing Punctuations
  - Translating the ASR Lattice
- Experiments
- Conclusions
Parameters

- We didn’t use discriminative training method to train the parameters. We adjust the parameters by hand.
- We didn’t readjust the parameters according to the develop sets provided in this evaluation. We simply used an empirical setting, with which our decoder achieved a good performance in translating the test set from the 2005 China’s National 863 MT Evaluation.
The parameter settings for our system

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Corresponding Features</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_1$</td>
<td>$p(\tilde{e}</td>
<td>\tilde{c})$</td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>$p(\tilde{c}</td>
<td>\tilde{e})$</td>
</tr>
<tr>
<td>$\lambda_3$</td>
<td>$lex(\tilde{e}</td>
<td>\tilde{c})$</td>
</tr>
<tr>
<td>$\lambda_4$</td>
<td>$lex(\tilde{c}</td>
<td>\tilde{e})$</td>
</tr>
<tr>
<td>$\lambda_5$</td>
<td>$lm(e_1')$</td>
<td>0.13</td>
</tr>
<tr>
<td>$\lambda_6$</td>
<td>$I$</td>
<td>0.48</td>
</tr>
<tr>
<td>$\lambda_7$</td>
<td>-$J$</td>
<td>0.48</td>
</tr>
</tbody>
</table>
Outline

- Overview
- Training
- System
  - Translation Model
  - Parameters
  - Decoder
  - Dealing with the Unknown Words
  - Recovering the Missing Punctuations
  - Translating the ASR Lattice
- Experiments
- Conclusions
Decoder

Source

Splitting the source

Sentence list

Loop

Translation list

Merging the translations

Target

Sentence

Preprocess

A monotone DP-based decoder

Postprocess

Translation

Chinese Language model

English Language model

Bilingual phrases

neon

Source

Splitting the source

Sentence list

Loop

Translation list

Merging the translations

Target

Sentence

Preprocess

A monotone DP-based decoder

Postprocess

Translation

Chinese Language model

English Language model

Bilingual phrases

neon
Decoder (Cont.)

- We used the monotone search in the decoding, similar to (Zens, 2002).
- Dynamic programming recursion:

\[ Q(0,\$) = 1 \]

\[ Q(j, e) = \max_{0 \leq j' < j, \tilde{e}, \tilde{e}'} \left\{ Q(j', e') + \sum_{m=1}^{M} \lambda_m \cdot h_m(\tilde{e}, c_{j'+1}^j) \right\} \]

\[ Q(J + 1,\$) = \max_{e'} \{ Q(J, e') + p(\$, e') \} \]
Outline

- Overview
- Training
- System
  - Translation Model
  - Parameters
  - Decoder
  - **Dealing with the Unknown Words**
  - Recovering the Missing Punctuations
  - Translating the ASR Lattice
- Experiments
- Conclusions
Dealing with the Unknown Words

- No special translation models for named entities are used. Named entities are translated in the same way as other unknown words.
- Unknown words were translated in two steps:
  - Firstly, we will look up a dictionary containing more than 100,000 Chinese words for the word.
  - If no translations are found in the first step, the word will then be translated using a rule-based Chinese-English translation system.
- All the 63 unknown words in the test data for the Cleaned Transcripts task in this evaluation are translated into English.
Outline

- Overview
- Training
- System
  - Translation Model
  - Parameters
  - Decoder
  - Dealing with the Unknown Words
  - Recovering the Missing Punctuations
  - Translating the ASR Lattice
- Experiments
- Conclusions
Recovering the Missing Punctuations

- There are no punctuations in the Chinese sentences.
- The missing of punctuations can have an adverse effect on the translation quality, so we developed a preprocessing model to recover the missing punctuations.
Recovering the Missing Punctuations (Cont.)

- **Two ways to do with the missing punctuations**
  - Method 1: To remove punctuations from the Chinese part of the training set, then train the model using the training set, and then translate the sentences without punctuations directly.
  - Method 2: To recover the punctuations from the input sentences, then translate the result sentences using a model trained from normal training set.

- **Experiments and results on develop set 4**

<table>
<thead>
<tr>
<th>Method</th>
<th>bleu-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>0.1936</td>
</tr>
<tr>
<td>Method 2</td>
<td>0.2139</td>
</tr>
</tbody>
</table>
Recovering the Missing Punctuations (Cont.)

- Given a Chinese sentence with $N$ words, $w_1 w_2 \ldots w_N$, we may construct a directed graph with $2N+1$ levels.
Recovering the Missing Punctuations (Cont.)

- Given such a graph, the problem of punctuation recovering could be looked on as a problem of searching the optimal path from the node in $level_0$ to the node in $level_{2N}$.

- In this problem, a path is said to be better than the other one if the language model score for it is larger than that for the latter.

- We then used the Viterbi algorithm to solve the search problem.
Outline

- Overview
- Training
- System
  - Translation Model
  - Parameters
  - Decoder
  - Dealing with the Unknown Words
  - Recovering the Missing Punctuations
  - Translating the ASR Lattice
- Experiments
- Conclusions
Translating the ASR Lattice

In the task of translating the ASR lattice, three types of test data were given:

- word lattice
- the 20-best results generated from ASR lattice
- the 1-best result generated from ASR lattice

A possibly better way is to regenerate the 1-best result based on Chinese language model from word lattice and then to translate it.
Translating the ASR Lattice (Cont.)

- We used a simpler approximate way
  - We first used our system to translate all the 20-best results and got 20 translations for each corresponding sentence.
  - Then we used the English language model to choose the best translation for each sentence.
Outline

- Overview
- Training
- System
  - Translation Model
  - Parameters
  - Decoder
  - Dealing with the Unknown Words
  - Recovering the Missing Punctuations
  - Translating the ASR Lattice
- Experiments
- Conclusions
Experiments

- The data we used

<table>
<thead>
<tr>
<th>Purposes</th>
<th>Corpus</th>
<th>statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilingual Phrase</td>
<td>Training set from IWSLT 2006</td>
<td>39,952 sentence pairs</td>
</tr>
<tr>
<td></td>
<td>Training set from the 2005 China’s National</td>
<td>152,049 sentence pairs</td>
</tr>
<tr>
<td></td>
<td>863 MT Evaluation</td>
<td></td>
</tr>
<tr>
<td>English Language Model</td>
<td>English part of the training set from the</td>
<td>7.4M words</td>
</tr>
<tr>
<td></td>
<td>2005 China’s National 863 MT Evaluation</td>
<td></td>
</tr>
<tr>
<td>Chinese Language Model</td>
<td>Chinese part of the training set from</td>
<td>350K Chinese words</td>
</tr>
<tr>
<td></td>
<td>IWSLT 2006</td>
<td></td>
</tr>
<tr>
<td>Chinese Reader (Duzhe) Corpus</td>
<td></td>
<td>7.9M Chinese words</td>
</tr>
</tbody>
</table>
Experiments (Cont.)

- The use of additional data did help improving the performance of our system on the develop sets.

- Influence of the additional bitexts (bleu-4)

<table>
<thead>
<tr>
<th></th>
<th>Training without additional bitexts</th>
<th>Training with additional bitexts</th>
</tr>
</thead>
<tbody>
<tr>
<td>develop set 1</td>
<td>0.3305</td>
<td>0.3922</td>
</tr>
<tr>
<td>develop set 2</td>
<td>0.3652</td>
<td>0.4349</td>
</tr>
<tr>
<td>develop set 3</td>
<td>0.4319</td>
<td>0.4823</td>
</tr>
<tr>
<td>develop set 4</td>
<td>0.1869</td>
<td>0.2139</td>
</tr>
</tbody>
</table>
Experiments (Cont.)

Scores of our system in IWSLT 2006

<table>
<thead>
<tr>
<th></th>
<th>official (with case + punctuation)</th>
<th>additional (without case + punctuation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE spontaneous speech ASR output</td>
<td>0.1505</td>
<td>0.1623</td>
</tr>
<tr>
<td>CE read speech ASR output</td>
<td>0.1579</td>
<td>0.1718</td>
</tr>
<tr>
<td>Correct Recognition Result</td>
<td>0.1976</td>
<td>0.2162</td>
</tr>
</tbody>
</table>
Some lessons

The scores on Correct Recognition Result are significantly **higher** than those on ASR output. This may result from the influence of the ASR errors. And the other reason may be the simple method we used to translate ASR lattice.
Experiments (Cont.)

- Some lessons (Cont.)
  - The scores on CE read speech ASR output are slightly higher than those on CE spontaneous speech ASR output. This indicates that the ASR system used to give the ASR output may be cleverer at the read speech data than at the spontaneous speech data.
Experiments (Cont.)

- Some lessons (Cont.)
  - The additional scores are higher than the official scores. This indicates that post-editing models such as truecasing or punctuation correction may help improving the translation quality. We will integrate such models in the future.
Outline

- Overview
- Training
- System
  - Translation Model
  - Parameters
  - Decoder
  - Dealing with the Unknown Words
  - Recovering the Missing Punctuations
  - Translating the ASR Lattice
- Experiments
- Conclusions
Conclusions

- We describe the system which participated in the 2006 IWSLT Speech Translation Evaluation of Institute of Artificial Intelligence, Xiamen University.
- It is a rather crude phrase-based SMT baseline, for example, without even considering phrase reordering.
- More improvements are underway.
References


References (Cont.)

- Koehn, Philipp, Axelrod, Amittai, Mayne, Alexandra Birch, Callison-Burch, Chris, Osborne, Miles and Talbot, David, "Edinburgh system description for the 2005 iwsit speech translation evaluation", *Proceeding of International Workshop on Spoken Language Translation*, Pittsburgh, PA, 2005
- He, Zhongjun, Liu, Yang, Xiong, Deyi, Hou, Hongxu and Liu, Qun, "ICT System Description for the 2006 TCSTAR Run #2 SLT Evaluation", *Proceeding of the TCSTAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, 2006, pp. 63-68.
Thanks

This work was supported by the National Natural Science Foundation of China (Grant No. 60573189 and Grant No. 60373080), National 863 High-tech Program (Grant No. 2004AA117010-06) and Key Research Project of Fujian Province (Grant No. 2006H0038)