CASIA Statistical Machine Translation
System for IWSLT 2008

Chengqing Zong
NLPR, Institute of Automation
Chinese Academy of Sciences
cqzong@nlpr.ia.ac.cn
Outline

- Our tasks
- System overview
- Technical modules
  - Preprocessing
  - Multiple translation engines
  - System combination
  - Rescoring
  - Post-processing
- Experiments
- Conclusions
Our tasks

We participated in:

1. Challenge task for Chinese-English
2. Challenge task for English-Chinese
3. BTEC task for Chinese-English.
System overview

- Using multiple translation engines
- Rescore the combination results to get the final translation outputs
Technical modules

- Preprocessing
  - Chinese
    - Chinese word segmentation
    - Transforming the Sexagesimal to Binary Converter (SBC) to Decimal to Binary Converter (DBC)
  - English
    - Tokenization of the English words - separates the punctuations with the English words;
    - Transforming the uppercase into lowercase.
Technical modules

- Phrase-based translation engines modeled in log-linear model

\[ e^* = \arg \max_e \sum_{m=1}^{M} \lambda_m h_m(e, f) \]

- Phrase translation probability;
- Lexical phrase translation probability;
- Inversed phrase translation probability;
- Inversed lexical phrase translation probability;
- English language model based on 3-gram;
- English sentence length penalty;
- Chinese phrase count penalty.
We use three phrase-based SMT:

- In-home developed phrase-based decoder (baseline)
- Moses decoder
- Bandore: A sentence type based reordering decoder
Preprocessing for PB SMT engine

- SVM is employed to divide the source (Chinese) sentences into three types, different types of sentences are reordered using different models

- Three types:
  - Special interrogative sentences
  - Other interrogative sentences
  - Non-question sentences
Technical modules- Bandore

- Architecture

- C1: special interrogative sentences
- C2: other interrogative sentences
- C3: non-question sentences
Special interrogative sentences

There is a fixed question phrase at the end of Chinese sentence, which is moved to the first position in the English translation. (We call the question phrase as Special Question Phrase)

你 想 要 什么样 的 座位 ？

What kind of seats do you like ?
Phrase-ahead reordering model moves the SQP to the frontal position in Chinese sentence

Two problems:
- Identification of SQP
- What position should SQP be moved to
Special words in SQP

Some Chinese words indicate the sentence is a special interrogative sentence

Close set: 什么 (what)、哪 (where)、多 (多长、多久) (how long)、怎 (how)、谁 (who, whom, whose)、几 (how many)、为什么 (why)、何 (when)
Definition of SQP:

- The syntactic component containing a special word in the close set

Identification:

- Use a shallow parsing toolkit (FlexCrf) (http://flexCRF.sourceforge.net)
Where should SQP be moved to?

- Three possible positions:
  - The beginning of the sentence
  - After the rightmost punctuation before the SQP
  - After a regular phrase such as “请问 (May I ask)” and “你 知道 (Do you know)”
这道菜怎么样？How about this dish?

你好，去海滩怎么走？Hello, how can I get to the beach?

你知道到那里需要多长时间？Do you know how long it takes us to there?
If we have known the SQP, S becomes $S^0$ SQP $S^1$, where $S^0$ is the left part of the sentence before SQP, and $S^1$ is the right part of the sentence after SQP. Therefore, we have learned the reordering templates from bilingual corpus to find the right position in $S^0$ where SQP will be moved to.
Other interrogative sentences

- Some specific Chinese words like “会、能、可以” are simply translated into “Can …”, “Do …” or “May …” at the beginning of the English sentence.

- This case is easy to process. So, we treat it as the no-question sentences
Non-question sentences

Some phrases are usually moved back during translation.

Three types of Chinese phrases are usually moved after the verb phrase in English sentence: (1) Prepositional phrase (PP), (2) Temporal phrase, and (3) Spatial phrase (SP).

我钱包在地铁里被偷了。

My wallet was stolen in the subway.
For the other interrogative sentences and non-question sentences, the phrase-back reordering model has been designed to move some phrases to the back positions.

- Two problems:
  - Identification of PP, TP, SP and VP
  - Reordering rules
Technical modules- Bandore

- Identification
  Use a shallow parsing toolkit
  (http://flexCRF.sourceforge.net)

- Reordering rules
  - Maximum entropy model is employed to decide whether a PP, TP or SP is moved back after VP
We develop a probabilistic reordering model to alleviate the impact of the errors caused by the parser when recognizing PPs, TPs, SPs and VPs. The form of phrase-back reordering rules:

$$A: \quad A_1 X A_2 \Rightarrow \begin{cases} A_1 X A_2 & \text{straight} \\ X A_2 A_1 & \text{inverted} \end{cases}$$

$$A_1 \in \{PP, TP, SP\}, \quad A_2 \in \{VP, FVP\}$$

$X$ is any phrases between $A_1$ and $A_2$. 
A Maximum Entropy Model is trained from bilingual spoken language corpus to determine whether $A_1$ should be moved after $A_2$:

$$P(O \mid A) = \frac{\exp\left(\sum_i \lambda_i h_i(O, A)\right)}{\sum_O \exp\left(\sum_i \lambda_i h_i(O, A)\right)}$$

$O \in \{\text{straight, inverted}\}$, $h_i(O, A)$ is a feature, and $\lambda_i$ is the weight.

The features include the leftmost, rightmost, and the POSs of $A_1$ and $A_2$. 
Technical modules

Other translation engines:

- Two formal syntax-based SMT engines:
  - HPB: A hierarchical phrase-based model
  - MEBTG: A maximum entropy-based reordering model
- A linguistically syntax-based SMT:
  - SAMT: A syntax-augmented machine translation decoder
System combination

- We implement system combination on \( N \)-Best list from multiple translation engines.
System combination

Find a hypothesis as the alignment reference with the minimum Bayesian risk

Align all the hypotheses against the alignment reference and forms a consensus alignment

Merge the similar words being aligned together at the same position and assign each word an alignment score based on a simple voting scheme. It thus forms a confusion network.

The final translation is found by the confusion network decoding with the language model feature and word penalty introduced.
Rescoring

- Use global feature functions to score the new n-best list
  - Direct and inverse IBM model 1 and model 3
  - 2, 4, 5-gram target language model
  - 3, 4, 5-gram target pos language model
  - Bi-word language model
  - Length ratio between source and target sentence
  - Question feature
  - Frequency of its n-gram ($n=1, 2, 3, 4$) within n-best translations
  - n-gram posterior probabilities within n-best translations.
  - Sentence length posterior probabilities.
Post-processing

The post-processing for the output results mainly includes:

- Case restoration in English words
- Recombination the separated punctuations with its left closest English words
- Segmenting the Chinese output into characters
Experiments

Corpus

- Besides the training data provided by IWSLT 2008, we collected all the data from the website of IWSLT2008.
- We extract the bilingual data which are highly correlative with the training data of each track.
- We also filter some development sentences and their reference sentences from all the released development data of the track as our development data according to the similarity calculation.
Experiments

* The detailed statistics of our corpus for development set

<table>
<thead>
<tr>
<th>Track</th>
<th>Data</th>
<th>Sen.</th>
<th>Running words</th>
<th>Voc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT</td>
<td>Train set</td>
<td>Chi 324,626</td>
<td>2.4M</td>
<td>11,214</td>
</tr>
<tr>
<td>CE</td>
<td></td>
<td>Eng 324,626</td>
<td>2.57M</td>
<td>9,488</td>
</tr>
<tr>
<td>CRR</td>
<td>Dev set</td>
<td>Chi 534</td>
<td>3.163</td>
<td>649</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Eng 3,204</td>
<td>22,861</td>
<td>1,132</td>
</tr>
<tr>
<td>CT</td>
<td>Train set</td>
<td>Chi 311,438</td>
<td>2.28M</td>
<td>11,113</td>
</tr>
<tr>
<td>EC</td>
<td></td>
<td>Eng 311,438</td>
<td>2.42M</td>
<td>9,370</td>
</tr>
<tr>
<td>CRR</td>
<td>Dev set</td>
<td>Chi 2275</td>
<td>15,266</td>
<td>797</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Eng 325</td>
<td>2,061</td>
<td>404</td>
</tr>
<tr>
<td>BTEC</td>
<td>Train set</td>
<td>Chi 321,770</td>
<td>2.38M</td>
<td>11,202</td>
</tr>
<tr>
<td>CCE</td>
<td></td>
<td>Eng 321,770</td>
<td>2.51M</td>
<td>9,493</td>
</tr>
<tr>
<td>CRR</td>
<td>Dev set</td>
<td>Chi 764</td>
<td>4,899</td>
<td>910</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Eng 4,584</td>
<td>34,310</td>
<td>1,536</td>
</tr>
</tbody>
</table>
Experiments

- **ASR translation**
  - We first translate the ASR $n$-best list.
  - For our experiments the value $n=5$ is used
  - We pass the translation results into our combination module and rescore all the translation hypotheses
  - With the feature functions of translation hypotheses plus the features of ASR
Experiments

- Results of development set for CT_CE track

<table>
<thead>
<tr>
<th></th>
<th>CRR</th>
<th>ASR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>NIST</td>
</tr>
<tr>
<td>PB</td>
<td>0.4505</td>
<td>7.4649</td>
</tr>
<tr>
<td>MOSES</td>
<td>0.5048</td>
<td>7.9175</td>
</tr>
<tr>
<td>Bandore</td>
<td>0.5033</td>
<td>8.0267</td>
</tr>
<tr>
<td>MEBTG</td>
<td>0.4571</td>
<td>7.6887</td>
</tr>
<tr>
<td>HPB</td>
<td>0.4412</td>
<td>6.8600</td>
</tr>
<tr>
<td>COM</td>
<td>0.5109</td>
<td>8.1780</td>
</tr>
<tr>
<td>Rescore</td>
<td>0.5741</td>
<td>8.3162</td>
</tr>
</tbody>
</table>
Experiments

- Results of development set for BTEC_CE track

<table>
<thead>
<tr>
<th></th>
<th>CRR</th>
<th>ASR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>NIST</td>
</tr>
<tr>
<td>PB</td>
<td>0.4659</td>
<td>7.9333</td>
</tr>
<tr>
<td>MOSES</td>
<td><strong>0.5100</strong></td>
<td>8.0298</td>
</tr>
<tr>
<td>Bandore</td>
<td><strong>0.5127</strong></td>
<td>8.3513</td>
</tr>
<tr>
<td>MEBTG</td>
<td>0.4717</td>
<td>7.8045</td>
</tr>
<tr>
<td>HPB</td>
<td>0.4764</td>
<td>6.5603</td>
</tr>
<tr>
<td>COM</td>
<td><strong>0.5308</strong></td>
<td>8.5689</td>
</tr>
<tr>
<td>Rescore</td>
<td><strong>0.6100</strong></td>
<td>8.7823</td>
</tr>
</tbody>
</table>
Experiments

- Results of development set for CT_EC track

<table>
<thead>
<tr>
<th></th>
<th>CRR</th>
<th>ASR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>NIST</td>
</tr>
<tr>
<td>PB</td>
<td>0.4385</td>
<td>7.0469</td>
</tr>
<tr>
<td>MEBTG</td>
<td>0.4399</td>
<td>7.5303</td>
</tr>
<tr>
<td>MOSES</td>
<td>0.4522</td>
<td>7.3626</td>
</tr>
<tr>
<td>HPB</td>
<td>0.4298</td>
<td>7.0914</td>
</tr>
<tr>
<td>COM</td>
<td>0.4555</td>
<td>7.6200</td>
</tr>
<tr>
<td>Rescore</td>
<td>0.5242</td>
<td>7.7361</td>
</tr>
</tbody>
</table>
Experiments

- Engines for combination on development set

<table>
<thead>
<tr>
<th></th>
<th>CT_CE</th>
<th></th>
<th></th>
<th>CT_EC</th>
<th></th>
<th></th>
<th>BTEC_CE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CRR</td>
<td>ASR</td>
<td>CRR</td>
<td>ASR</td>
<td>CRR</td>
<td>ASR</td>
<td></td>
</tr>
<tr>
<td>PB</td>
<td></td>
<td></td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>MOSES</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Bandore</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>MEBTG</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>HPB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td>√</td>
<td></td>
</tr>
</tbody>
</table>
Experiments

Results of test set for each track

- Con1: our system combination
- Con2: the rescoring module
- Primary: we RE-rescore “Con1” and “Con2” by using the feature of the prior probability of the length-ratio of source sentence to target sentence.
## Experiments

<table>
<thead>
<tr>
<th>Track</th>
<th>System</th>
<th>CRR</th>
<th>ASR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BLEU</td>
<td>NIST</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.4844</td>
<td>7.5859</td>
</tr>
<tr>
<td>CE</td>
<td>Primary</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Con1</td>
<td>0.4803</td>
<td>7.4277</td>
</tr>
<tr>
<td></td>
<td>Con2</td>
<td>0.4767</td>
<td>7.4237</td>
</tr>
<tr>
<td>CE</td>
<td>Primary</td>
<td>0.5122</td>
<td>7.3513</td>
</tr>
<tr>
<td></td>
<td>Con1</td>
<td>0.4968</td>
<td>7.1525</td>
</tr>
<tr>
<td></td>
<td>Con2</td>
<td>0.4817</td>
<td>6.7254</td>
</tr>
<tr>
<td>CE</td>
<td>Primary</td>
<td>0.5077</td>
<td>8.5389</td>
</tr>
<tr>
<td></td>
<td>Con1</td>
<td>0.4842</td>
<td>8.4094</td>
</tr>
<tr>
<td></td>
<td>Con2</td>
<td>0.5162</td>
<td>8.2884</td>
</tr>
</tbody>
</table>

The table above shows the performance metrics for different tracks and systems, with CRR and ASR scores for BLEU and NIST.
Experiments

- The best performance relatively compared with PB decoder among the scores on development set.

<table>
<thead>
<tr>
<th>System</th>
<th>Compared with PB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bandore</td>
<td>11.72%</td>
</tr>
<tr>
<td>MEBTG</td>
<td>5.03%</td>
</tr>
<tr>
<td>HPB</td>
<td>4.45%</td>
</tr>
</tbody>
</table>
Conclusions

- Our system combines the output results of multiple machine translation engines and by using some global features we rescore the combination results to get the final translation outputs.
Conclusions

- In all the translation engines, Moses has a performance with considerable robustness.
- Bandore has an outstanding performance among the three engines:
  - It uses Moses as its decoder.
  - The reordering model of Bandore aims at the spoken language. It has an effective ability to translation in the domain of IWSLT.
Thanks

谢谢！