The CASIA SMT System

Chengqing Zong

National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences

cqzong@nlpr.ia.ac.cn
Outline

1. Description of CASIA SMT System
2. Experiments
3. Conclusion
1. Description of CASIA SMT System

- Foundation of SMT Research in CASIA
  - Chinese word Segmentation and POS tagging
  - Syntactic parsing
  - Large scale corpora
  - Machine translation in some small domains:
    - based on interlingua (IF)
    - example-based system
    - word-based on IBM model1 and IBM model2
    - hybrid translation system
1. Description of CASIA SMT System

- The CASIA SMT system
  - Chinese-to-English
  - in the domain of BTEC
  - phrase-based SMT
  - size of the training corpus is much larger than given 20,000 sentence pairs
1. Description of CASIA SMT System

- The phrase-based SMT model

\[ p(e \mid c) = p_T(c \mid e)^{l_t} \times p_L(e)^{l_i} \times p_D(e, c)^{l_d} \]

Where, \( p_T(c \mid e) \) is the translation model;
\( p_L(e) \) is the target language model;
\( p_D(e, c) \) is the distortion model and
\[ p_D(e, c) = \lambda |a_i - b_{i-1} - 1| \]
1. Description of CASIA SMT System

Components: Translation model, language model, decoder
1. Description of CASIA SMT System

◆ **Phrase Translation Model**

Four methods for learning phrase translation:

• IBM word-based model 4 [P. Brown, 1993]

• Integrated segmentation and phrase alignment (ISA) [Zhang, 2003]

• HMM word alignment model [Vogel, 1996]

• Giza++ toolkit
1. Description of CASIA SMT System

- Phrase translation for “我要买”

<table>
<thead>
<tr>
<th>English Phrase</th>
<th>$\phi (f/e)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I want to buy</td>
<td>0.386</td>
</tr>
<tr>
<td>I would like to buy</td>
<td>0.234</td>
</tr>
<tr>
<td>I will buy</td>
<td>0.119</td>
</tr>
<tr>
<td>I wanna buy</td>
<td>0.108</td>
</tr>
<tr>
<td>I wan to get</td>
<td>0.101</td>
</tr>
<tr>
<td>......</td>
<td>......</td>
</tr>
</tbody>
</table>
1. Description of CASIA SMT System

- **Language Model**
  - Standard Technique: Trigram Model
    - multiplication of trigram probabilities
  - Using the available SRI language modeling toolkit.

http://www.speech.sri.com/projects/srilm/
1. Description of CASIA SMT System

◆ Beam-search for decoding

- Look up possible phrase translations [Koehn, 2003]

- many different ways to segment words into phrases

- many different ways to translate each phrase

<table>
<thead>
<tr>
<th>中国</th>
<th>与</th>
<th>北朝鲜</th>
<th>有</th>
<th>外交</th>
<th>关系</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>with</td>
<td>North Korea</td>
<td>has</td>
<td>diplomatic</td>
<td>relationships</td>
</tr>
</tbody>
</table>

- China has the diplomatic relationships with North Korea
Decoding: hypothesis expansion

- Build the English by hypothesis expansion
  - from the left to right
  - search space exponential with sentence length
  - reduction by pruning weak hypothesis (keep top 200)

China has the diplomatic relationships with North Korea
Decoding: Stack Organization

*The hypotheses are stored in different stacks*

- bin label: number of Chinese words covered (One, Two)
- insert the Functional Words (of, the, for.....) bins.

F-Words must expand to Non F-Words Next
Decoding: search the best translation

Considering there are many auxiliary words and mood words in the Chinese sentence, and these words sometimes don’t have the corresponding words in the English sentence. So, the system does not require the all words in the Chinese sentence to be translated.

Select the candidate sentences generated after $L - a$ Chinese words have been translated.

$L$ is the length of the input Chinese sentence; $a$ is an integer.
Decoding: search the best translation

For example:

$L = 7$ words

input: 我 想 预订 一 个 单人间
(I want to reserve a single room)

if $a=1$, input = 我 想 预订 一
I would like to book one
I want to reserve one

if $a=2$, input = 我 想 预订
I want to book one
I would like to reserve one

... ...
1. Description of CASIA SMT System

- Pre-processing of numerals using rule-based approach
  - Arabic numbers: telephone No., room No. etc
  - Number written in Chinese, such as “一百 one hundred”
  - Ordinal numbers:
  - Dates
  - Combination of the different expressions
1. Description of CASIA SMT System

- After pre-processing, the numerals are replaced with the specific marks (variables).

- The phrase translations become the templates with variables, e.g.,

  \[ \text{X 个单人间} \rightarrow \text{X single rooms} \]
1. Description of CASIA SMT System

- In our experiment, about 5% extracted phrases contain variable.

- The performance of the system has been improved about 8.2% using pre-processing of the numerals.
2. Experiments

- Experiment-1
- Comparison of the different methods for learning phrase translations using 100,000 sentence pairs

<table>
<thead>
<tr>
<th>Methods</th>
<th>Num of Phrases</th>
<th>BLEU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM4</td>
<td>194,802</td>
<td>0.2683</td>
</tr>
<tr>
<td>ISA</td>
<td>187,011</td>
<td>0.2751</td>
</tr>
<tr>
<td>HMM</td>
<td>278,770</td>
<td>0.2637</td>
</tr>
<tr>
<td>Giza++</td>
<td>695,486</td>
<td>0.2882</td>
</tr>
</tbody>
</table>
2. Experiments

- **Experiment-2**
  - Comparison of the different searching algorithms using 100,000 sentence pairs

**Marks:**
- *M* means word-based translation model;
- +NF0 means the Functional-zero words are not applied;
- +F0 means the Functional-zero words are applied;
- +NUM means the numerals are pre-processed;
- +BACK1 stands for our decoder;
- +BACK2 stands for Koehn’s decoder.
2. Experiments

<table>
<thead>
<tr>
<th>Methods</th>
<th>Bleu (4-gram)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M+NF0+BACK2</td>
<td>0.1833</td>
</tr>
<tr>
<td>M+NF0+BACK1</td>
<td>0.1919</td>
</tr>
<tr>
<td>M+F0+BACK2</td>
<td>0.2372</td>
</tr>
<tr>
<td>M+F0+BACK1</td>
<td>0.2663</td>
</tr>
<tr>
<td>Giza++ +NF0+BACK2</td>
<td>0.2730</td>
</tr>
<tr>
<td>Giza++ +NF0+BACK1</td>
<td>0.2864</td>
</tr>
<tr>
<td>Giza++ +F0+BACK2</td>
<td>0.2763</td>
</tr>
<tr>
<td>Giza++ +F0+BACK1</td>
<td>0.2882</td>
</tr>
<tr>
<td>EDM+NF0+BACK1</td>
<td>0.1978</td>
</tr>
<tr>
<td>EDM+F0+BACK1</td>
<td>0.2683</td>
</tr>
<tr>
<td>Giza++ +F0+BACK1+NUM</td>
<td>0.3177</td>
</tr>
</tbody>
</table>
2. Experiments

Experiment-3

- Comparison of the number of translation options and decoding time using 900,000 sentence pairs

<table>
<thead>
<tr>
<th>Methods</th>
<th>Bleu (4-gram)</th>
<th>Decoding time</th>
</tr>
</thead>
<tbody>
<tr>
<td>G+F0+BACK1</td>
<td>0.3418</td>
<td>2H6Min</td>
</tr>
<tr>
<td>G+F0+BACK1_Top100</td>
<td>0.3452</td>
<td>40Min</td>
</tr>
<tr>
<td>G+F0+BACK1_Top150</td>
<td>0.3446</td>
<td>54Min</td>
</tr>
<tr>
<td>G+F0+BACK1_Top200</td>
<td>0.3423</td>
<td>64Min</td>
</tr>
<tr>
<td>G+F0+BACK1_Top50</td>
<td>0.3366</td>
<td>23Min</td>
</tr>
</tbody>
</table>
2. Experiments

- IWSLT2005 evaluation
- Training Corpus
  - 1M sentence pairs in the specific domain of C-Star, including BTEC corpus and CJK corpus and CASIA corpus
  - 500K sentence pairs in the general domain (news) from Chinese LDC
2. Experiments

- Perplexity of source language (Chinese)
  - Use SRILM tool
  - Results:
    - counting all input tokens: 41.2084
    - excluding end-of-sentence tags: 69.3387
2. Experiments

Results from IWSTL’2005

<table>
<thead>
<tr>
<th>Track (C-E)</th>
<th>Data condition</th>
<th>Bleu4</th>
<th>NIST</th>
<th>Meteor</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual transcription</td>
<td>unrestricted</td>
<td>0.5279</td>
<td>10.2499</td>
<td>0.7214</td>
<td>0.4160</td>
<td>0.3366</td>
</tr>
<tr>
<td>ASR Output</td>
<td>unrestricted</td>
<td>0.3845</td>
<td>8.0406</td>
<td>0.5802</td>
<td>0.5788</td>
<td>0.4770</td>
</tr>
</tbody>
</table>

Top 20 ASR results
3. Conclusion

In the phrase-based CASIA Chinese-to-English SMT system:

• the numerals are pre-processed, so the phrase translations contain variables

• in decoding, some functional words are added and the output candidates are selected from the generated English sentences after $L - a$ Chinese words have been translated.
3. Conclusion

We are inspired by the pre-processing of the numerals, and we believe that the named entity translation will increase the system performance.

Next step:  - named entity translation
- phrase translation extraction
谢谢！

Thanks!