The ICT Statistical Machine Translation Systems for IWSLT 2007

Zhongjun He, Haitao Mi, Yang Liu, Devi Xiong, Weihua Luo, Yun Huang, Zhixiang Ren, Yajuan Lu, Qun Liu

Institute of Computing Technology
Chinese Academy of Sciences
2007.09.15– 2007.08.16
Outline

- Overview
- MT Systems
  - Bruin
  - Confucius
  - Lynx
- Official Evaluation
- Discussion
- Summary
Introduction of Our Group

- Multilingual Interaction Technology Laboratory, Institute of Computing Technology, Chinese Academy Sciences
- Long history for working on MT
  - Rule-based
  - Example-based
- Focus on SMT from 2004
- Website: http://mtgroup.ict.ac.cn/
People Working on SMT at ICT

- **Staffs**
  - Qun Liu (Researcher)
  - Yajuan Lu (Associate Researcher)
  - Yang Liu (Associate Researcher)
  - Weihua Luo (Assistant Researcher)

- **PhD Students**
  - Zhongjun He
  - Haitao Mi
  - Jinsong Su
  - Yang Feng

- **Master Students**
  - Yun Huang
  - Wenbin Jiang
  - Zhixiang Ren
  - ...
IWSLT 2007 Evaluation

Chinese-English transcript translation task
Systems for IWSLT 2007 Evaluation

MT Systems:
- *Bruin* (formally syntax-based)
- *Confucius* (extended phrase-based)
- *Lynx* (linguistically syntax-based)
Outline

- Overview
- MT Systems
  - Bruin
  - Lynx
  - Confucius
- Official Evaluation
- Discussion
- Summary
**Bruin**

- Bruin is a formally syntax-based system
- MaxEnt Reordering Model build on BTG rules

![Diagram showing straight and inverted models with target and source nodes]

- Regard reordering as a binary classification
  - Building a MaxEnt-based classifier
  - Using boundary words instead of whole phrases as features for the classifier
**Features**

- Source and target boundary words (lexical feature)
- Combinations of boundary words (collocation feature)

\[
h_i(o, b^1, b^2) = \begin{cases} 1, & b^1.t_1 = E_1, o = O \\ 0, & \text{otherwise} \end{cases} \]

\[
h_j(o, b^1, b^2) = \begin{cases} 1, & b^1.t_1 = E_1, b^2.t_1 = E_2, o = O \\ 0, & \text{otherwise} \end{cases} \]
Training and Decoding

- Training the model
  - Learning reordering examples from bilingual word-aligned corpus
  - Generating features from reordering examples
  - Training a MaxEnt model on the features

- Decoding
  - CKY algorithm

For details, see Xiong et al., ACL2006
Confucius

- An extended phrase-based system
- Log-linear model
- Monotone decoding
- We try a phrase-based similarity model, in which a translation for a certain source phrase can be applied for other similar phrases
Phrase-based Similarity Model

全省 出口 总值 的 25.5%

Find the most similar phrase pair

全市 出口 总值 的 半数

half of the entire city's export volume
Phrase-based Similarity Model

全省 出口 总值 的 25.5%

Compare

全市 出口 总值 的 半数

half of the entire city's export volume
Phrase-based Similarity Model

全省 出口 总值 的 25.5%

Replace

全省 出口 总值 的 25.5%

25.5% of the entire province's export volume
**Lynx**

- A linguistically syntax-based system
- Based on tree-to-string alignment template (TAT), which map the source language tree to target language string
- Log-linear Model
Translation Process: Parsing

中国 的 经济 发展

down arrow

parsing

NP

DNP  NP

NP  DEG  NN  NN

NR  的  经济  发展

中国
Translation Process: Detachment
Translation Process: Production

- "的" to "of"
- "中国" to "China"
- "经济" to "economic"
- "发展" to "development"
Translation Process: Combination

economic development of China
Training and Decoding

- **Training**
  - Extract TATs from word-aligned, source side parsed bilingual corpus using bottom-up strategy
  - Impose several restrictions to decrease the magnitude
- **Decoding**
  - bottom-up beam search
- For details, see [Liu et al., ACL2006](#)
Outline

● Overview
● MT Systems
  ● Bruin
  ● Lynx
  ● Confucius
● Official Evaluation
● Discussion
● Summary
Toolkits Used

- Word alignment
  - GIZA++ plus “grow-diag-final” refinement method
- Language model
  - SRILM
- Chinese parser
  - Deyi Xiong’s
    - A lexicalized PCFG model trained on PennTree bank
- Chinese word segmentation
  - ICTCLAS
Preprocessing and Postprocessing

- **Preprocessing**
  - Chinese word segmentation
  - Rule-based translations of numbers, dates and Chinese names
  - Chinese sentences Parsing (for Lynx only)

- **Postprocessing**
  - Remove unknown words
  - Capitalize the first word of each sentence
# Training data

<table>
<thead>
<tr>
<th>Names</th>
<th>Description</th>
<th>Sentence Pairs</th>
<th>Chinese Words</th>
<th>English Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>IWSLT2007</td>
<td>Training data provided by IWSLT 2007</td>
<td>39,943</td>
<td>354k</td>
<td>378k</td>
</tr>
<tr>
<td>LDC2002L27</td>
<td>Chinese-English Translation Lexicon Version 3.0</td>
<td>79,369</td>
<td>79k</td>
<td>123k</td>
</tr>
<tr>
<td>2004-863-008</td>
<td>Dialog corpus from ChineseLDC</td>
<td>51,694</td>
<td>486k</td>
<td>509k</td>
</tr>
<tr>
<td>CLDC-LAC-2003-004</td>
<td>Chinese-English Sentence aligned Bilingual Corpus from ChineseLDC</td>
<td>199,702</td>
<td>2.7M</td>
<td>3.1M</td>
</tr>
<tr>
<td>CLDC-LAC-2003-006</td>
<td>Chinese-English Sentence aligned Bilingual Corpus from ChineseLDC</td>
<td>299,952</td>
<td>4.5M</td>
<td>4.7M</td>
</tr>
</tbody>
</table>

Training Data List
## Development and test set

<table>
<thead>
<tr>
<th></th>
<th>Chinese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>IWSLT’06-dev Sentences</td>
<td>489</td>
<td></td>
</tr>
<tr>
<td>Running Words</td>
<td>5983</td>
<td>45720</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>1139</td>
<td>2150</td>
</tr>
<tr>
<td>IWSLT’06-tst Sentences</td>
<td>500</td>
<td></td>
</tr>
<tr>
<td>Running Words</td>
<td>6359</td>
<td>51227</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>1331</td>
<td>2346</td>
</tr>
<tr>
<td>IWSLT’07-tst Sentences</td>
<td>489</td>
<td></td>
</tr>
<tr>
<td>Running Words</td>
<td>3297</td>
<td>22574</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>879</td>
<td>1527</td>
</tr>
</tbody>
</table>

Corpus statistics of the IWSLT 2006 and 2007 development and test set
Results on IWSLT 2006 development set and test set

<table>
<thead>
<tr>
<th>Condition</th>
<th>System Name</th>
<th>IWSLT'06-dev</th>
<th>IWSLT'06-tst</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small Data</td>
<td>Bruin</td>
<td>0.1756</td>
<td>0.1731</td>
</tr>
<tr>
<td></td>
<td>Confucius</td>
<td>0.1724</td>
<td>0.1700</td>
</tr>
<tr>
<td></td>
<td>Lynx</td>
<td>0.1681</td>
<td>0.1667</td>
</tr>
<tr>
<td>Large Data</td>
<td>Bruin</td>
<td>0.2114</td>
<td>0.2283</td>
</tr>
<tr>
<td></td>
<td>Confucius</td>
<td>0.2115</td>
<td>0.2042</td>
</tr>
<tr>
<td></td>
<td>Lynx</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Small data: The training data released by the IWSLT 2007
Large data: All the training data
Results on IWSLT 2007 test set

<table>
<thead>
<tr>
<th>System Name</th>
<th>IWSLT’07-tst</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bruin</td>
<td>0.3750</td>
</tr>
<tr>
<td>Confucius</td>
<td>0.2802</td>
</tr>
<tr>
<td>Lynx</td>
<td>0.1777</td>
</tr>
</tbody>
</table>
Outline

- Overview
- MT Systems
  - Bruin
  - Lynx
  - Confucius
- Official Evaluation
- Discussion
- Summary
Discussion

**Lynx(0.1777)**

**Training Corpus:**

- Training data:
  - About 39k sentence pairs **dialogs** data
    - Provided by IWSLT 2007
  - About 5M sentence pairs **newswire** data
    - Released by LDC

- Domain is quite different
  - **Newswire** vs. **Dialogs**

- **Newswire** data is too large
Discussion

Lynx (0.1777)

Parser:

• Trained on Penn Chinese Treebank
• Domain is quite different too
  – Newswire vs. Dialogs
• Parsing error (low performance of parser)
• Lynx decoder
  – Only depends on the 1-best parsing tree
Discussion

Models:
- Bruin (0.3750)
- Confucius (0.2802)
Discussion

Models:
- Bruin (0.3750)
  - MaxEnt based reordering model
  - Long distance word reordering
- Confucius (0.2802)
  - Monotone search

2007 test set       (2006 test set)
- 6.7 words/sent    (12.7 words/sent)
  - Bruin will do better
  - Punctuation marks  (no )
  - More positive reordering information
  - Bruin will do better
Discussion

**Models:**

- **Bruin (0.3750)**
  - MaxEnt based reordering model
  - Long distance word reordering
- **Confucius (0.2802)**
  - Monotone search

<table>
<thead>
<tr>
<th></th>
<th>2006 tst</th>
<th>2007 tst</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bruin</td>
<td>0.2283</td>
<td>0.3750</td>
</tr>
<tr>
<td>Confucius</td>
<td>0.2042</td>
<td>0.2802</td>
</tr>
</tbody>
</table>

- **2007 test set** (2006 test set)
Discussion

Models:

- **Bruin** (0.3750)
  - MaxEnt based reordering model
  - Long distance word reordering
- **Confucius** (0.2802)
  - Monotone search

- **2007 test set** (2006 test set)
  - 6.7 words/sent  (12.7 words/sent)
    - Bruin will do better

<table>
<thead>
<tr>
<th></th>
<th>2006 tst</th>
<th>2007 tst</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bruin</td>
<td>0.2283</td>
<td>0.3750</td>
</tr>
<tr>
<td>Confucius</td>
<td>0.2042</td>
<td>0.2802</td>
</tr>
</tbody>
</table>
Discussion

Models:

- **Bruin (0.3750)**
  - MaxEnt based reordering model
  - Long distance word reordering

- **Confucius (0.2802)**
  - Monotone search

2007 test set (2006 test set)

- 6.7 words/sent (12.7 words/sent)
  - Bruin will do better
- Punctuation marks (no)
  - More positive reordering information
  - Bruin will do better
Results on IWSLT 2007 test set

<table>
<thead>
<tr>
<th>System Name</th>
<th>IWSLT’07-tst</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bruin</td>
<td>0.3750</td>
</tr>
<tr>
<td>Confucius</td>
<td>0.2802</td>
</tr>
<tr>
<td>Lynx</td>
<td>0.1777</td>
</tr>
</tbody>
</table>
Outline

- Overview
- Systems
  - Bruin
  - Lynx
  - Confucius
- Official Evaluation
- Discussion
- Summary
Summary

MT
- 3 systems based on different translation models:
  - MaxEnt BTG Model
  - TAT model
  - Phrase-based Similarity Model

Future Work
- More new model
- System combination
References


Thanks!