The TCH Machine Translation System for IWSLT 2008

Haifeng Wang
Toshiba (China) R&D Center
Oct. 20, 2008
Outline

• Overview

• Modules
  – Dictionary, segmentation, alignment, NE
  – LM, Punctuation and case restoration

• Tasks
  – CE: CT, BTEC
  – EC: CT
  – CS: BTEC
  – CES: PIVOT

• Evaluation Results

• Summary
Introduction

• Tasks
  – BTEC tasks: BTEC_CE, BTEC_CS
  – Challenge tasks: CT_CE, CT_EC
  – Pivot task: PIVOT_CES

• Methods
  – SMT
  – RBMT
  – Pivot SMT
  – Combination
  – Module improvement

• Resources
  – Supplied resources provided for each data track
  – Other publicly available resources
MT Methods

• SMT
  – Phrase-based SMT: Moses

• Pivot SMT
  (Wu and Wang, ACL 2007)
  – Phrase translation probability
  – Lexical weight

• RBMT
  – Publicly available software: Dr. eye

• Combination of RBMT and SMT
  (Hu, Wang and Wu, EMNLP 2007)
  – Using SMT system as the main MT system
  – Using RBMT system to produce synthetic bilingual corpus
  – The SMT system is trained using both real and synthetic corpus

• Translation selection
  – 5-gram LM method (Chen et al. IWSLT 2006)
  – Target sentence average length method
Modules

- Dictionary
- Chinese Word Segmentation
- Word Alignment
- Named Entity Translation
- Language Model
- Punctuation Restoration
- Case Restoration
Bilingual Dictionary

- Existed bilingual dictionary
  - General dictionary
    - LDC Chinese-English translation lexicon
  - NE dictionary
    - LDC Chinese <-> English Name Entity Lists
    - Person names and location names

- Dictionary extracted from corpus
  - Automatically extracted from in-domain corpus
    - Bidirectional word alignment
    - Filtering
      - Translation probability
      - Co-occurring frequency
  - Check
Chinese Monolingual Dictionary

• Dictionaries
  – General dictionary
    • Extracted from LDC Chinese-English lexicon
  – NE dictionary
    • Extracted from LDC Chinese-English NE list
  – In-domain dictionary
    • Extracted from in-domain corpus

• Word granularity
  – Tune the word unit referring to its translation in target language
    • Word
    • Multi-word expression
Chinese Word Segmentation

• Initial experiments
  – Segmentation ambiguity in domain-specific spoken language is not serious

• Segmentation method
  – Forward maximum-matching
    • Basic segmentation method
  – Back one character method
    • To indentify ambiguous fragment
  – Ambiguous fragments database
    • For disambiguation

• Word normalization
  – To deal with data sparseness
  – Extract a synonym list from translation dictionary and corpus
  – Only used when Chinese is source language
Word Alignment

- **Alignment algorithm**
  - Bidirectional word alignment using IBM models
  - Keep links in the intersection set
  - Keep links occurring in bilingual dictionaries
  - Delete links conflicting with the links in the final alignment set
  - Keep remained links
  - Different alignment heuristics
    - Grow-diag (CE, CS), grow (EC), grow-diag-final (ES)

- **Resources**
  - Bilingual corpus
  - General dictionary
  - Domain-specific dictionary
Named Entity Translation

• NE recognition and translation

<table>
<thead>
<tr>
<th>Method</th>
<th>Digit</th>
<th>Date</th>
<th>Time</th>
<th>Person name</th>
<th>Location name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule</td>
<td>Rule</td>
<td>Rule</td>
<td>Rule</td>
<td>Dictionary</td>
<td>Dictionary</td>
</tr>
</tbody>
</table>

• NE processing in SMT
  – Training
    • Replace NEs in the training data with NE tags
    • Train model on the data with NE tags
  – Translating
    • Replace NEs in the input sentence with NE tags
    • Translate
    • Restore the NE tags with their translations
Language Model

• In-domain corpus
  – Target language part of the provided corpus for a given track

• Out-of-domain corpus
  – Publicly available corpus
  – Selection
    • Perplexity
    • Sentence
    • Using the in-domain LM

• Interpolation
  – Linear interpolation using SRILM
  – Weight tuned on development sets
Punctuation Restoration

• Restore punctuation in source language
• English
  – Hidden-ngram (SRILM toolkit)
  – Rules
    • By hand
    • Based on some keywords, e.g. a sentence begin with “could”
• Chinese
  – Maximum entropy model
  – 2 steps
    • Position determination
    • Punctuation determination
  – Features
    • Words around a boundary
    • Words at the beginning or end of a sub-sentence
Case Restoration

• Restore case in target language
  – English
  – Spanish

• Method
  – recaser
    • In the training scripts of Moses
    • As a MT problem
    • Trained on the corpus with case information
  – Lexicon based post-processing
    • To process English words that should be capitalized
      Such as proper nouns
    • The lexicon is extracted from some available resources
      Such as training text in respective tasks, HIT corpus, Tanaka corpus
Tasks

• Five tasks
  – Chinese-English
    • Challenge task (CT_CE)
    • BTEC task (BTEC_CE)
  – English-Chinese
    • Challenge task (CT_EC)
  – Chinese-Spanish
    • BTEC task (BTEC_CS)
  – Chinese-English-Spanish
    • Pivot task (PIVOT_CES)

• Input
  – Spontaneous speech (SS)
  – Read speech (RS)
  – Correct recognition result (CRR)
Chinese-English Tasks – Data

• Dictionary

<table>
<thead>
<tr>
<th>Type</th>
<th>General</th>
<th>Domain</th>
<th>NE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>LDC2002L27</td>
<td>Extracted from In-domain corpus</td>
<td>LDC2005T34</td>
</tr>
<tr>
<td>Number</td>
<td>54,170</td>
<td>38,620</td>
<td>47692</td>
</tr>
</tbody>
</table>

• Training Corpus

<table>
<thead>
<tr>
<th>Corpus</th>
<th>BTEC</th>
<th>HIT</th>
<th>CLDC</th>
<th>Tanaka</th>
</tr>
</thead>
<tbody>
<tr>
<td># sentence pairs</td>
<td>19,972</td>
<td>80,868</td>
<td>200,732</td>
<td>149,207</td>
</tr>
<tr>
<td># source words</td>
<td>177,168</td>
<td>802,454</td>
<td>2,113,534</td>
<td>-</td>
</tr>
<tr>
<td># target words</td>
<td>182,627</td>
<td>822,508</td>
<td>2,096,731</td>
<td>1,351,645</td>
</tr>
</tbody>
</table>

– Selection and preprocessing

• Development set
  – devset1, devset2, devset4

• Test set
## Chinese-English Tasks – Experimental Results

**Results (Case sensitive BLEU score, CRR input)**

<table>
<thead>
<tr>
<th>Method</th>
<th>devset3</th>
<th>devset5</th>
<th>devset6</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBMT</td>
<td>0.4253</td>
<td>0.2020</td>
<td>0.2086</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.5186</td>
<td>0.2013</td>
<td>0.2807</td>
</tr>
<tr>
<td>Our segmenter</td>
<td>0.5425</td>
<td>0.2047</td>
<td>0.3029</td>
</tr>
<tr>
<td>+HIT</td>
<td>0.5697</td>
<td>0.2323</td>
<td>0.3416</td>
</tr>
<tr>
<td>+Dic</td>
<td>0.5819</td>
<td>0.2375</td>
<td>0.3456</td>
</tr>
<tr>
<td>+NE</td>
<td>0.5838</td>
<td>0.2396</td>
<td>0.3537</td>
</tr>
<tr>
<td>+CLDC</td>
<td>0.5891</td>
<td>0.2445</td>
<td>0.3554</td>
</tr>
<tr>
<td>+RBMT</td>
<td>0.6091</td>
<td>0.2536</td>
<td>0.3570</td>
</tr>
<tr>
<td>+LM Inter.</td>
<td>0.6223</td>
<td>0.2516</td>
<td>0.3823</td>
</tr>
</tbody>
</table>

**Translation selection**

- **Mert**
  - Default: default in Moses
  - Mert1: best on devset5
  - Mert2: Stable

<table>
<thead>
<tr>
<th>Method</th>
<th>devset3</th>
<th>devset5</th>
<th>devset6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>0.5927</td>
<td>0.2547</td>
<td>0.3453</td>
</tr>
<tr>
<td>Mert1</td>
<td>0.6061</td>
<td>0.2679</td>
<td>0.3837</td>
</tr>
<tr>
<td>Mert2</td>
<td>0.6274</td>
<td>0.2551</td>
<td>0.3863</td>
</tr>
<tr>
<td>Select</td>
<td>0.6260</td>
<td>0.2627</td>
<td>0.3882</td>
</tr>
</tbody>
</table>
English-Chinese Tasks – Data

• Dictionary
  – General dictionary, domain dictionary, NE dictionary
    (Same as CE tasks)

• Training Corpus

<table>
<thead>
<tr>
<th>Corpus</th>
<th>BTEC</th>
<th>HIT</th>
</tr>
</thead>
<tbody>
<tr>
<td># sentence pairs</td>
<td>19,972</td>
<td>89,318</td>
</tr>
<tr>
<td># source words</td>
<td>189,041</td>
<td>945,010</td>
</tr>
<tr>
<td># target words</td>
<td>178,339</td>
<td>914,121</td>
</tr>
</tbody>
</table>

  – Selection
  – Preprocessing
    • English abbreviation restoration
    • Without Chinese word normalization

• Development and test set
  – devset, devset3
  – No MERT
## English-Chinese Tasks – Experimental Results

### Results

<table>
<thead>
<tr>
<th></th>
<th>devset3</th>
<th>devset</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBMT</td>
<td>0.4362</td>
<td>0.4425</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.4455</td>
<td>0.4511</td>
</tr>
<tr>
<td><strong>Our segmenter</strong></td>
<td>0.4528</td>
<td>0.4564</td>
</tr>
<tr>
<td>+Dic</td>
<td>0.4551</td>
<td>0.4684</td>
</tr>
<tr>
<td>+NE</td>
<td>0.4558</td>
<td>0.4773</td>
</tr>
<tr>
<td>+HIT</td>
<td>0.4830</td>
<td>0.5325</td>
</tr>
<tr>
<td>+RBMT</td>
<td>0.5131</td>
<td>0.5426</td>
</tr>
<tr>
<td>+Select</td>
<td>0.5133</td>
<td>0.5551</td>
</tr>
</tbody>
</table>

### Translation selection
- 2 Candidates
  - Without RBMT
  - With RBMT
  - Selection metric: LM
Chinese-Spanish Tasks

• Training Corpus
  – BTEC data provided for this task
  – Preprocessing similar as CE task
• Dictionary
  – Extracted from the training corpus (9990 entries)
• Test set
  – Devset3
• Post-processing
  – Rule-based, such as question mark "?" and "¿"
• Experimental Results

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Our segmenter</th>
<th>+dic</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>0.3596</td>
<td>0.3726</td>
<td>0.3839</td>
</tr>
</tbody>
</table>
Chinese-English-Spanish – Data

• Dictionary
  – LDC CE dictionary
  – CE dictionary extracted from BTEC and HIT CE corpus (39010)
  – ES dictionary extracted from BTEC and Europarl ES corpus (10426)

• Training Corpus

<table>
<thead>
<tr>
<th>Corpus</th>
<th>BTEC CE</th>
<th>HIT CE</th>
<th>BTEC ES</th>
<th>Europarl ES</th>
<th>Tanaka</th>
</tr>
</thead>
<tbody>
<tr>
<td># sentence pairs</td>
<td>20,000</td>
<td>80,868</td>
<td>19,972</td>
<td>400,000</td>
<td>149,207</td>
</tr>
<tr>
<td># source words</td>
<td>164,957</td>
<td>802,454</td>
<td>182,627</td>
<td>8,485,253</td>
<td>-</td>
</tr>
<tr>
<td># target words</td>
<td>182,793</td>
<td>822,508</td>
<td>185,527</td>
<td>8,219,380</td>
<td>1,351,645</td>
</tr>
</tbody>
</table>

  – Selection and preprocessing

• Test set
  – devset3
Chinese-English-Spanish – Experimental Results

• Results

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>+dic+HIT+Europarl</th>
<th>+RBMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pivot model</td>
<td>0.2791</td>
<td>0.3616</td>
<td>0.4136</td>
</tr>
<tr>
<td>Transfer model</td>
<td>0.3243</td>
<td>0.4139</td>
<td>0.4423</td>
</tr>
<tr>
<td>Trans. selection</td>
<td>-</td>
<td>-</td>
<td>0.4510</td>
</tr>
</tbody>
</table>

• RBMT
  – Translate the English part of ES corpus into Chinese -> synthetic CE corpus
  – Synthetic CE corpus is used in pivot and transfer model

• Transfer model is better than pivot model
  – CE translation is quite good (0.6024)
  – English and Spanish are more similar than Chinese and Spanish
  – pivot model contains much more noise than the transfer model

• Translation selection
  – Selection metric: length
## IWSLT 2008 Evaluation Results

<table>
<thead>
<tr>
<th>Method</th>
<th>(Bleu + Meteor)/2</th>
<th>Bleu</th>
<th>Meteor</th>
<th>Human Eval.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CT_EC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS</td>
<td>0.5647</td>
<td>0.4818</td>
<td>0.6476</td>
<td>0.3906</td>
</tr>
<tr>
<td>CRR</td>
<td>0.6566</td>
<td>0.5912</td>
<td>0.7219</td>
<td>-</td>
</tr>
<tr>
<td><strong>CT_CE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS</td>
<td>0.5257</td>
<td>0.4166</td>
<td>0.6347</td>
<td>0.4516</td>
</tr>
<tr>
<td>CRR</td>
<td>0.5909</td>
<td>0.4980</td>
<td>0.6837</td>
<td>-</td>
</tr>
<tr>
<td><strong>BTEC_CE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RS</td>
<td>0.5358</td>
<td>0.4474</td>
<td>0.6241</td>
<td>0.4730</td>
</tr>
<tr>
<td>CRR</td>
<td>0.5887</td>
<td>0.5085</td>
<td>0.6688</td>
<td>-</td>
</tr>
<tr>
<td><strong>BTEC_CS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RS</td>
<td>0.3273</td>
<td>0.3218</td>
<td>0.3328</td>
<td>0.4316</td>
</tr>
<tr>
<td>CRR</td>
<td>0.3597</td>
<td>0.3582</td>
<td>0.3611</td>
<td>-</td>
</tr>
<tr>
<td><strong>PIVOT_CES</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RS</td>
<td>0.3620</td>
<td>0.3657</td>
<td>0.3583</td>
<td>0.4624</td>
</tr>
<tr>
<td>CRR</td>
<td>0.4044</td>
<td>0.4157</td>
<td>0.3931</td>
<td>-</td>
</tr>
</tbody>
</table>
Summary

• Tasks
  – BTEC_CE, BTEC_CS, CT_CE, CT_EC, PIVOT_CES

• Resources
  – Supplied resources provided for each data track
  – Other Publicly available resources

• Methods
  – Adaptation of Chinese word segmentation
  – Word alignment refinement using dictionary and various heuristics
  – Named entities translation
  – Additional corpus (In-domain, Out-of-domain)
  – Combination of SMT and RBMT
  – Translation selection
Thanks!