The CNGL-DCU-Prompsit Translation Systems for WMT13

Raphael Rubino†, Antonio Toral†, Santiago Cortés Vaillo*, Jun Xie§, Xiaofeng Wu†, Stephen Doherty†, Qun Liu‡

†NCLT/CNGL, Dublin City University, Ireland  ‡ICT, Chinese Academy of Sciences, China  *Prompsit Language Engineering, Spain

{rrubino, atoral, xfwu, qliu}@computing.dcu.ie  stephen.doherty@dcu.ie

The CNGL-DCU-Prompsit Translation Systems for WMT13

Raphael Rubino†, Antonio Toral†, Santiago Cortés Vaillo*, Jun Xie§, Xiaofeng Wu†, Stephen Doherty†, Qun Liu‡

†NCLT/CNGL, Dublin City University, Ireland  ‡ICT, Chinese Academy of Sciences, China  *Prompsit Language Engineering, Spain

{rrubino, atoral, xfwu, qliu}@computing.dcu.ie  stephen.doherty@dcu.ie

Introduction

- Goal: investigate the use of linguistic information to select parallel data in a state-of-the-art PB-SMT pipeline

Setting

- Software: Moses 1.0, IRSTLM 5.80.01
- LMs: 5-gram, improved Kneser-Ney, built for each monolingual corpus, interpolated with WMT11 test as devset
- Decoding parameters: cube pruning, MBR and monotone at punctuation reordering

Training data: Europarl (EU), News Commentary (NC), United Nations (UN) and Common Crawl (CC)

LM data: EN, NC, News 2007–2012 and target side of CC, UN and 10^6 French–English corpus (only for English as target language)

Data Selection

- Perplexity-based approach to filter parallel data (Axelrod et al., 2011)
- Measure perplexity not only on word forms but also using different types of linguistic information (lemmas and named entities)
- Subsets with lowest perplexity for 2 thresholds (e.g. 5% with forms, 14% with lemmas+nes for EU)

<table>
<thead>
<tr>
<th>corpus</th>
<th>size</th>
<th>forms</th>
<th>forms+nes</th>
<th>lemmas</th>
<th>lemmas+nes+nes</th>
</tr>
</thead>
<tbody>
<tr>
<td>EU</td>
<td>5%</td>
<td>957.9</td>
<td>987.2</td>
<td>974.3</td>
<td>1005.5</td>
</tr>
<tr>
<td></td>
<td>14%</td>
<td>1108.4</td>
<td>1095.7</td>
<td>1111.6</td>
<td>1057.7</td>
</tr>
<tr>
<td>UN</td>
<td>2%</td>
<td>877.1</td>
<td>969.6</td>
<td>866.6</td>
<td>982.2</td>
</tr>
<tr>
<td></td>
<td>12%</td>
<td>1203.2</td>
<td>1130.9</td>
<td>1183.8</td>
<td>1131.6</td>
</tr>
<tr>
<td>CC</td>
<td>50%</td>
<td>573.9</td>
<td>547.2</td>
<td>574.5</td>
<td>546.4</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td>560.1</td>
<td>560.1</td>
<td>560.1</td>
<td>560.1</td>
</tr>
</tbody>
</table>

Results

- Evaluate systems using selected subsets of corpora (small: 5% EU, 2% UN, 50% CC, big: 14% EU, 12% UN, 100% CC) for English–Spanish

<table>
<thead>
<tr>
<th>System #sent</th>
<th>BLEU</th>
<th>BLEUcased</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>small</td>
<td>1.4M</td>
<td>33.12</td>
<td>32.05</td>
</tr>
<tr>
<td>big</td>
<td>3.8M</td>
<td>33.49</td>
<td>32.43</td>
</tr>
</tbody>
</table>

French–English

Introduction

- Goal: to evaluate the gain of adding small in-domain parallel data into a PB-SMT built on a sub-sample of the out-of-domain parallel data.
- Two phrase tables are built: one on News-Commentary, one on an out-of-domain 3.7M pairs sub-sample (Modified Moore-Lewis).
- Both EN-FR and FR-EN are tested.

Language Models

- Modified Kneser-Ney 5-gram LMs (with SRILM).
- In-Domain LMs: News-Commentary and News-Crawl.
- LDC Gigaword LMs: one LM per news source.
- LMs are first interpolated within groups based on WMT dev sets.

Translation Models

- MGiza++ and Moses v1.0.
- Decoder configuration: 50 target phrases limit, reordering constraint.

Results

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>BLEUdev</td>
<td>26.9</td>
</tr>
<tr>
<td>BLEU</td>
<td>27.0</td>
</tr>
<tr>
<td>BLEUcased</td>
<td>26.1</td>
</tr>
<tr>
<td>TER</td>
<td>62.9</td>
</tr>
<tr>
<td>BLEU and TER scores obtained by our systems: BLEUdev is the score obtained on the development set given by MERT, while BLEU, BLEUcased and TER are obtained on the test set given by the submission website.</td>
<td></td>
</tr>
</tbody>
</table>

German–English

Introduction

- Goal: to explore the performance of three baseline systems, as well as a system combination approach.
- Baseline systems: a PB-SMT (Moses), a HPB-SMT (Moses) and a HPB-SMT (Cdec)

Preprocessing and Corpus

- Normalization, truecasing, filter with standard clean scripts. Split the German compound words with WordSplitter.
- All data for training: WMT 2011 and 2012 test sets for development.
- All the data, including GigawordV5 for LM.
- The method we used to train Language Model is provided in French-English subtask.

Three Baseline Systems

- The motivation of choosing Hierarchical Models is to address the German-English’s long reorder problem.
- We want to test the performance of Cdec and Moses HPB-SMT and choose the best.
- Moses PB-SMT is used as our benchmark

<table>
<thead>
<tr>
<th>Development Test</th>
<th>BLEU Moses</th>
<th>HPB Moses</th>
<th>Cdec</th>
<th>Combination</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>22.0</td>
<td>24.1</td>
<td>22.5</td>
<td>24.4</td>
</tr>
</tbody>
</table>

Results

- A word-level combination strategy.
- 5 best hypothesis as the alignment reference for the Confusion Network (CN).
- HMM to choose the backbone.
- Tune with Simple-Simplex algorithm. The parameters for system weights are set equal.

System Combination

- The CNGL-DCU-Prompsit Translation Systems for WMT13

Raphael Rubino†, Antonio Toral†, Santiago Cortés Vaillo*, Jun Xie§, Xiaofeng Wu†, Stephen Doherty†, Qun Liu‡

†NCLT/CNGL, Dublin City University, Ireland  ‡ICT, Chinese Academy of Sciences, China  *Prompsit Language Engineering, Spain

{rrubino, atoral, xfwu, qliu}@computing.dcu.ie  stephen.doherty@dcu.ie