Creating a Bilingual Dictionary using Wikipedia

MSc Thesis Defense Presentation

Angelina Ivanova

Advisor: RNDr. Daniel Zeman, Ph.D.

September 6, 2011
Overview

➢ Introduction
➢ Motivation
➢ Experiments
➢ Conclusions
➢ Future Work
Goals

- **Develop** dictionaries from the Wikipedia link structure
- **Evaluate** quality and examine the content
- **Apply** the dictionary to statistical machine translation task
• Bilingual dictionaries are important resources for NLP

• Manual development is expensive

• Internet-resources provide terminology which is not present in traditional dictionaries
**Interlanguage links** - links from an article to a presumably equivalent, article in another language.

[[cs:Tygr džunglový]]
[[en:Tiger]]
Redirect pages - pages without article body that provide:

- equivalent names for an entity
  
  **Federation of Czechoslovakia** -> **Czechoslovakia**

- spelling resolution
  
  **Microsoft’s** -> **Microsoft**

- spelling correction
  
  **Wikipedia** -> **Wikipedia**
Development of Wiki-dictionaries

Redirect page

CSFR

Entity page

Czechoslovakia

Entity page

Чехословакия

Redirect page

Чехословацкая Социалистическая Республика

Redirect page

Entity page

Redirect page

Federation of Czechoslovakia

Interlanguage link

Czechoslovakia | Federation of Czechoslovakia | CSFR
Чехословакия | Чехословацкая Социалистическая Республика | ЧСФР
Named Entity Recognition

Heuristics (Bunescu and Pasca, 2006) + our additional heursitics

Results of the named entity recognition

88% named entities
12% common words

74% accuracy on the sample of size 100
False positives:

- All content words of a multi-word special term are capitalized
  
  Museum of Fine Arts
  Cloning Extinct Species

- Phrases contain named entities
  
  Pakistan at the 1948 Summer Olympics
  Saturn Award for Best Make-up

- Common words containing two or more capital letters
  
  WebConferencing
False negatives:

- Short articles for one-word named entities
  Muaná
- Pairs that occur only in Russian Wikipedia
Named entity classification

- Applied only to the titles of entity pages
- Types: PER, LOC, ORG, MISC
- Bootstrapping algorithm (Knopp, 2010)
- Final labels obtained by:
  - Translations of labels for En titles
  - Translations of labels for Ru titles
  - Intersection of labels from the labels for En and Ru titles
- Corrections using heuristics based on the comments in brackets
# Evaluation of named entity classification

<table>
<thead>
<tr>
<th>Intersection of En and Ru NE</th>
<th>NE labels translated from En</th>
<th>NE labels translated from Ru</th>
</tr>
</thead>
<tbody>
<tr>
<td># of entries in the sample</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td># of correctly classified NE</td>
<td>170</td>
<td>180</td>
</tr>
</tbody>
</table>
Sample data: English-Russian

Asahi, Mie
Асахи (Миэ)
LOC

Atul Dodiya/Atul dodiya
Додиа, Атул|Атул Додиа
PER

Bad Kohlgrub
Бад-Кольгруб
LOC

Banda people/Banda (CAR)/Banda (Sudan)
Банда (народ)
Sample data: English-German

*Beak* / *Beaks* / *Bird’s beak* / *Rhamphotheca*
*Schnabel* | *Oberschnabel* | *Unterschnabel*

*Calw (district)* / *Calw district* / *Landkreis Calw*
*Landkreis Calw* | *Kreis Calw*
*LOC*

*Christian Pander*
*Christian Pander* | *Funky Pee*
*PER*
Abbreviations (2204 entries)
Geographical names (1282 entries)
Names (630 entries)
Base (50695 entries)
Recall:

- Geographical names: **82.18%**
- Names: **75.88%**
- Abbreviations: **22.64%**
- Base: **7.42%**
Corpus Statistics

Statistics of the Wiki-dictionary on UMC corpus

Approx. **28%** of the training set – **0** translation pairs from the Wiki-dictionaries

Approx. **24.7%** of the training set - **1** translation pair from the Wiki-dictionaries
**BLEU Score for the Trained Models**

<table>
<thead>
<tr>
<th></th>
<th>BLEU score</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-gram</td>
<td>21.19</td>
</tr>
<tr>
<td><strong>4-gram</strong></td>
<td><strong>21.42</strong></td>
</tr>
<tr>
<td>5-gram</td>
<td>20.99</td>
</tr>
<tr>
<td>4-gram + additional data for LM</td>
<td>24.60</td>
</tr>
<tr>
<td><strong>5-gram + additional data for LM</strong></td>
<td><strong>24.76</strong></td>
</tr>
<tr>
<td>3-gram + Wiki-dict.</td>
<td>20.05</td>
</tr>
<tr>
<td><strong>4-gram + Wiki-dict.</strong></td>
<td><strong>20.42</strong></td>
</tr>
<tr>
<td>5-gram + Wiki-dict.</td>
<td>20.38</td>
</tr>
</tbody>
</table>
### Test Sets Statistics

<table>
<thead>
<tr>
<th></th>
<th>UMC test set</th>
<th>Wiki test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>numb. of sentences</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>tokens</td>
<td>19,019</td>
<td>22,498</td>
</tr>
<tr>
<td>vocabulary size</td>
<td>7,001</td>
<td>7,694</td>
</tr>
</tbody>
</table>
## Manual Ranking

<table>
<thead>
<tr>
<th></th>
<th>Model <em>w/t Wiki-dict</em> is ranked first</th>
<th>Model <em>with Wiki-dict</em> is ranked first</th>
<th>Translations are equally bad/good</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 sample from UMC test set</td>
<td>55</td>
<td>37</td>
<td>8</td>
</tr>
<tr>
<td>100 sample from Wiki-set</td>
<td>44</td>
<td>50</td>
<td>6</td>
</tr>
</tbody>
</table>
## SMT Experiments

### Out-of-Vocabulary Words Statistics

<table>
<thead>
<tr>
<th></th>
<th>w/t Wiki-dict.</th>
<th>with Wiki-dict</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong># of OOV</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>UMC test set</strong></td>
<td>934</td>
<td>906</td>
</tr>
<tr>
<td><strong>Wiki test set</strong></td>
<td>2,260</td>
<td>1,878</td>
</tr>
</tbody>
</table>
Translation with model trained without Wiki-dictionary:

after the death of фредди меркьюри remaining members groups, using records his votes, could provoke in 1995, the latest krylenko queen - made in heaven.

Translation with model trained with Wiki-dictionary:

after the death of freddie mercury remaining members of groups, through records his votes, managed to issue in 1995, the last queen made in heaven.
Experiment without comments in brackets (UMC set)

Damo *(philosopher)* | Damo *(Greece)*
Дамо *(философ)*

Results:

- BLEU: 20.89 vs 20.42
- OOV: 889 (929) vs 906 (945)
Conclusions

✔ Wiki-dictionaries and traditional dictionaries differ dramatically (7.42% recall, named entities, noun phrases)

✔ Wiki-dictionaries can cause drop of accuracy in MT experiment (due to domain shift)

✔ The methods can be applied to any language pair which is present in Wikipedia (e.g. English-German)
Future Work

- Evaluation on a parallel corpus from another domain
- Connection of the dictionary to the morphological analyzer
- Factored machine translation
- Improvement of named entity recognition and classification
- The impact of the comments in brackets and additional information
Thank you for your attention!

Questions?