LINA: Identifying Comparable Documents from Wikipedia

Emmanuel Morin$^1$ Amir Hazem$^2$ Elizaveta Loginova-Clouet$^1$ Florian Boudin$^1$

$^1$LINA - UMR CNRS 6241, Université de Nantes, France

$^2$LIUM - EA 4023, Université du Maine, France

BUCC-2015 Shared Task
Introduction

- How far can we go with a language agnostic model?
- We experiment with [Enright and Kondrak, 2007]'s parallel document identification method
- We adapt the method to the BUCC-2015 Shared task based on two assumptions:
  1. Source documents should be paired 1-to-1 with target documents
  2. We have access to comparable documents in several languages
Outline

Introduction

Method

Experiments

Summary
Method

- Fast parallel document identification [Enright and Kondrak, 2007]
  - Documents = bags of hapax words
  - Words = blank separated strings that are 4+ characters long
  - Given a document in language A, the document in language B that shares the largest number of words is considered as parallel

- Works very well for parallel documents
  - 99.96% accuracy on EUROPARL [Enright and Kondrak, 2007]
  - 80% precision on Wikipedia [Patry and Langlais, 2011]

- We use this approach as baseline for detecting comparable documents
Improvements using 1-to-1 alignments

- In baseline, document pairs are scored independently
  - Multiple source documents are paired to a same target document
  - \(\approx 60\%\) of English pages are paired with multiple pages in French or German

- We remove multiply assigned source documents using *pigeonhole* reasoning
  - From 60% to 11% of multiply assigned source documents

![Diagram representing document relationships](image-url)
Improvements using cross-lingual information

- Simple document weighting function → score ties
- We break the remaining score ties using a third language
  - From 11% to less than 4% of multiply assigned source documents
Outline

Introduction

Method

Experiments

Summary
Experimental settings

- We focus on the French-English and German-English pairs

- The following measures are considered relevant
  - Mean Average Precision (MAP)
  - Success (Succ.)
  - Precision at 5 (P@5)
## Results (FR $\rightarrow$ EN)

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Train</th>
<th></th>
<th></th>
<th>Test</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAP</td>
<td>Succ.</td>
<td>P@5</td>
<td>MAP</td>
<td>Succ.</td>
<td>P@5</td>
</tr>
<tr>
<td>baseline</td>
<td>31.4</td>
<td>28.0</td>
<td>7.4</td>
<td>32.9</td>
<td>30.0</td>
<td>7.5</td>
</tr>
<tr>
<td>+ pigeonhole</td>
<td>57.7</td>
<td>56.4</td>
<td>11.9</td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>+ cross-lingual</td>
<td>58.9</td>
<td>57.7</td>
<td>12.1</td>
<td>59.0</td>
<td>57.7</td>
<td>12.1</td>
</tr>
</tbody>
</table>
## Results (DE→EN)

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Train</th>
<th></th>
<th>Test</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAP</td>
<td>Succ.</td>
<td>P@5</td>
<td>MAP</td>
</tr>
<tr>
<td>baseline</td>
<td>28.7</td>
<td>24.9</td>
<td>6.9</td>
<td>29.0</td>
</tr>
<tr>
<td>+ pigeonhole</td>
<td>61.6</td>
<td>60.1</td>
<td>12.8</td>
<td>−</td>
</tr>
<tr>
<td>+ cross-lingual</td>
<td>62.3</td>
<td>60.9</td>
<td>12.8</td>
<td>62.2</td>
</tr>
</tbody>
</table>
Outline

Introduction

Method

Experiments

Summary
Summary

- Unsupervised, hapax words-based method
- Promising results, about 60% of success using pigeonhole reasoning
- Using a third language slightly improves the performance

Future work
- Finding the optimal alignment across the all languages
- Relaxing the hapax-words constraint
Thank you

florian.boudin@univ-nantes.fr