LINA: Identifying Comparable Documents from Wikipedia
Emmanuel Morin, Amir Hazem, Florian Boudin, Elizaveta Loginova Clouet

To cite this version:

HAL Id: hal-01185670
https://hal.archives-ouvertes.fr/hal-01185670
Submitted on 21 Aug 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
LINA: Identifying Comparable Documents from Wikipedia

Emmanuel Morin\textsuperscript{2} Amir Hazem\textsuperscript{1} Elizaveta Loginova-Clouet\textsuperscript{2} Florian Boudin\textsuperscript{2}

\textsuperscript{1}LIUM - EA 4023, Université du Maine, France
amir.hazem@lium.univ-lemans.fr

\textsuperscript{2}LINA - UMR CNRS 6241, Université de Nantes, France
\{elizaveta.loginova,florian.boudin,emmanuel.morin\}@univ-nantes.fr

Abstract

This paper describes the LINA system for the BUCC 2015 shared track. Following (Enright and Kondrak, 2007), our system identify comparable documents by collecting counts of hapax words. We extend this method by filtering out document pairs sharing target documents using pigeonhole reasoning and cross-lingual information.

1 Introduction

Parallel corpora, that is, collections of documents that are mutual translations, are used in many natural language processing applications, particularly for statistical machine translation. Building such resources is however exceedingly expensive, requiring highly skilled annotators or professional translators (Preiss, 2012). Comparable corpora, that are sets of texts in two or more languages without being translations of each other, are often considered as a solution for the lack of parallel corpora, and many techniques have been proposed to extract parallel sentences (Munteanu et al., 2004; Abdul-Rauf and Schwenk, 2009; Smith et al., 2010), or mine word translations (Fung, 1995; Rapp, 1999; Chiao and Zweigenbaum, 2002; Morin et al., 2007; Vulić and Moens, 2012).

Identifying comparable resources in a large amount of multilingual data remains a very challenging task. The purpose of the Building and Using Comparable Corpora (BUCC) 2015 shared task\textsuperscript{1} is to provide the first evaluation of existing approaches for identifying comparable resources. More precisely, given a large collection of Wikipedia pages in several languages, the task is to identify the most similar pages across languages.

\textsuperscript{1}https://comparable.limsi.fr/bucc2015/

In this paper, we describe the system that we developed for the BUCC 2015 shared track and show that a language agnostic approach can achieve promising results.

2 Proposed Method

The method we propose is based on (Enright and Kondrak, 2007)’s approach to parallel document identification. Documents are treated as bags of words, in which only blank separated strings that are at least four characters long and that appear only once in the document (hapax words) are indexed. Given a document in language A, the document in language B that share the largest number of these words is considered as parallel.

Although very simple, this approach was shown to perform very well in detecting parallel documents in Wikipedia (Patry and Langlais, 2011). The reason for this is that most hapax words are in practice proper nouns or numerical entities, which are often cognates. An example of hapax words extracted from a document is given in Table 1. We purposely keep urls and special characters, as these are useful clues for identifying translated Wikipedia pages.

| website major gaston links flutist marcel debost states sources college crunelle conservatoire principal rampal united currently recorded chastain competitions music | http://www.oberlin.edu/faculty/mdebost/under international flutists jean-pierre profile moyse french repertoire amazon lives external *http://www.amazon.com/michel-debost/dp/b000s9zsk0 known teaches conservatory school professor studied kathleen orchestre replaced michel |

Table 1: Example of indexed document as bag of hapax words (en-bacde.txt).
Here, we experiment with this approach for detecting near-parallel (comparable) documents. Following (Patry and Langlais, 2011), we first search for the potential source-target document pairs. To do so, we select for each document in the source language, the \( N = 20 \) documents in the target language that share the largest number of hapax words (hereafter baseline).

Scoring each pair of documents independently of other candidate pairs leads to several source documents being paired to a same target document. As indicated in Table 2, the percentage of English articles that are paired with multiple source documents is high (57.3% for French and 60.4% for German). To address this problem, we remove potential multiple source documents by keeping the document pairs with the highest number of shared words (hereafter pigeonhole). This strategy greatly reduces the number of multiply assigned source documents from roughly 60% to 10%. This in turn removes needlessly paired documents and greatly improves the effectiveness of the method.

![Figure 1: Example of the use of cross-lingual information to order multiple documents that received the same scores.](image)

**Table 2:** Percentage of English articles that are paired with multiple French or German articles on the training data.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>FR→EN</th>
<th>DE→EN</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>57.3</td>
<td>60.4</td>
</tr>
<tr>
<td>+ pigeonhole</td>
<td>10.7</td>
<td>10.8</td>
</tr>
<tr>
<td>+ cross-lingual</td>
<td>3.7</td>
<td>3.4</td>
</tr>
</tbody>
</table>

3 Experiments

3.1 Experimental settings

The BUCC 2015 shared task consists in returning for each Wikipedia page in a source language, up to five ranked suggestions to its linked page in English. Inter-language links, that is, links from a page in one language to an equivalent page in another language, are used to evaluate the effectiveness of the systems. Here, we only focus on the French-English and German-English pairs. Following the task guidelines, we use the following evaluation measures investigate the effectiveness of our method:

- **Mean Average Precision** (MAP). Average of precisions computed at the point of each correctly paired document in the ranked list of paired documents.
- **Success** (Succ.). Precision computed on the first returned paired document.
- **Precision at 5** (P@5). Precision computed on the 5 topmost paired documents.

3.2 Results

Results are presented in Table 3. Overall, we observe that the two strategies that filter out multiply assigned source documents improve the performance of the method. The largest part of the improvement comes from using pigeonhole reasoning. The use of cross-lingual information to
Table 3: Performance in terms of MAP, success (Succ.) and precision at 5 (P@5) of our model.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Train MAP</th>
<th>Train Succ.</th>
<th>Train P@5</th>
<th>Test MAP</th>
<th>Test Succ.</th>
<th>Test P@5</th>
</tr>
</thead>
<tbody>
<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>

break ties between the remaining multiply assigned source documents only gives a small improvement. We assume that the limited number of potential source-target document pairs we use in our experiments ($N = 20$) is a reason for this.

Interestingly, results are consistent across languages and datasets (test and train). Our best configuration, that is, with pigeonhole and cross-lingual, achieves nearly 60% of success for the first returned pair. Here we show that a simple and straightforward approach that requires no language-specific resources still yields some interesting results.

4 Discussion

In this paper we described the LINA system for the BUCC 2015 shared track. We proposed to extend (Enright and Kondrak, 2007)'s approach to parallel document identification by filtering out document pairs sharing target documents using pigeonhole reasoning and cross-lingual information. Experimental results show that our system identifies comparable documents with a precision of about 60%.

Scoring document pairs using the number of shared hapax words was first intended to be a baseline for comparison purposes. We tried a finer grained scoring approach relying on bilingual dictionaries and information retrieval weighting schemes. For reasonable computation time, we were unable to include low-frequency words in our system. Partial results were very low and we are still in the process of investigating the reasons for this.

Acknowledgments

This work is supported by the French National Research Agency under grant ANR-12-CORD-0020.

References


