LINA: Identifying Comparable Documents from Wikipedia

Emmanuel Morin, Amir Hazem, Florian Boudin, Elizaveta Loginova Clouet

LINA: Identifying Comparable Documents from Wikipedia

Emmanuel Morin[2] **Amir Hazem**[1] **Elizaveta Loginova-Clouet**[2] **Florian Boudin**[2]

[1] LIUM - EA 4023, Université du Maine, France
amir.hazem@lium.univ-lemans.fr

[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[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.

[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).

Proceedings of the Eighth Workshop on Building and Using Comparable Corpora, pages 88–91,
Beijing, China, July 30, 2015.

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.

Strategy	**FR→EN**	**DE→EN**
baseline	57.3	60.4
+ pigeonhole	10.7	10.8
+ cross-lingual	3.7	3.4

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

In an attempt to break the remaining score ties between document pairs, we further extend our model to exploit cross-lingual information. When multiple source documents are paired to a given English document with the same score, we use the paired documents in a third language to order them (hereafter *cross-lingual*). Here we make two assumptions that are valid for the BUCC 2015 shared Task: (1) we have access to comparable documents in a third language, and (2) source documents should be paired 1-to-1 with target documents.

An example of two French documents (doc_{fr} 1 and doc_{fr} 2) being paired to the same English document (doc_{en}) is given in Figure 1. We use the German document (doc_{de}) paired with doc_{en} and select the French document that shares the largest number of hapax words, which for this example is doc_{fr} 2. This strategy further reduces the number of multiply assigned source documents from 10% to less than 4%.

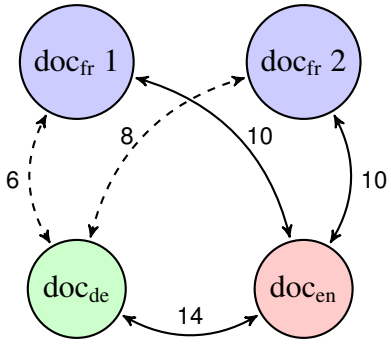

Figure 1: Example of the use of cross-lingual information to order multiple documents that received the same scores. The number of shared words are labelled on the edges.

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

	FR→EN						DE→EN					
	Train			Test			Train			Test		
Strategy	MAP	Succ.	P@5	MAP	Succ.	P@5	MAP	Succ.	P@5	MAP	Succ.	P@5
baseline	31.4	28.0	7.4	32.9	30.0	7.5	28.7	24.9	6.9	29.0	24.9	7.1
+ pigeonhole	57.7	56.4	11.9	–	–	–	61.6	60.1	12.8	–	–	–
+ cross-lingual	58.9	57.7	12.1	59.0	57.7	12.1	62.3	60.9	12.8	62.2	60.7	12.8

Table 3: Performance in terms of MAP, success (Succ.) and precision at 5 (P@5) of our model.

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

Sadaf Abdul-Rauf and Holger Schwenk. 2009. On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 16–23, Athens, Greece.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 2*, COLING '02, pages 1–5, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jessica Enright and Grzegorz Kondrak. 2007. A fast method for parallel document identification. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'07)*, pages 29–32, Rochester, New York, USA.

Pascale Fung. 1995. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In *Proceedings of the 3rd Annual Workshop on Very Large Corpora (VLC'95)*, pages 173–183, Cambridge, MA, USA.

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. 2007. Bilingual terminology mining - using brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 664–671, Prague, Czech Republic, June. Association for Computational Linguistics.

Dragos Stefan Munteanu, Alexander Fraser, and Daniel Marcu. 2004. Improved machine translation performance via parallel sentence extraction from comparable corpora. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 265–272, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Alexandre Patry and Philippe Langlais. 2011. Identifying parallel documents from a large bilingual collection of texts: Application to parallel article extraction in wikipedia. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora (BUCC'11)*, pages 87–95, Portland, Oregon, USA.

Judita Preiss. 2012. Identifying comparable corpora using lda. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 558–562, Montréal, Canada, June. Association for Computational Linguistics.

Reinhard Rapp. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 519–526, College Park, MD, USA.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 403–411, Los Angeles, California, June. Association for Computational Linguistics.

Ivan Vulić and Marie-Francine Moens. 2012. Detecting highly confident word translations from comparable corpora without any prior knowledge. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 449–459, Avignon, France, April. Association for Computational Linguistics.