

LINA: Identifying Comparable Documents from Wikipedia

Emmanuel Morin¹ Amir Hazem² Elizaveta Loginova-Clouet¹ Florian Boudin¹

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

²LIUM - 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

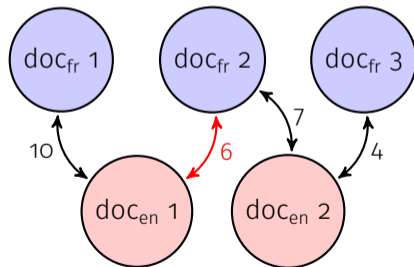
Summary

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

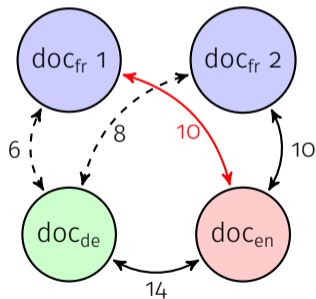
Improvements using 1-to-1 alignments

- ▶ 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



Improvements using cross-lingual information

- ▶ Simple document weighting function \rightarrow 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→EN)

Strategy	Train			Test		
	MAP	Succ.	P@5	MAP	Succ.	P@5
baseline	31.4	28.0	7.4	32.9	30.0	7.5
+ pigeonhole	57.7	56.4	11.9	—	—	—
+ cross-lingual	58.9	57.7	12.1	59.0	57.7	12.1

Results (DE→EN)

Strategy	Train			Test		
	MAP	Succ.	P@5	MAP	Succ.	P@5
baseline	28.7	24.9	6.9	29.0	24.9	7.1
+ pigeonhole	61.6	60.1	12.8	—	—	—
+ cross-lingual	62.3	60.9	12.8	62.2	60.7	12.8

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`

References I

 Enright, J. and Kondrak, G. (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.

 Patry, A. and Langlais, P. (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.