CLIR- and ontology-based approach for bilingual extraction of comparable documents

Manuela Yapomo, Gloria Corpas, Ruslan Mitkov

CLIR- and ontology-based approach for bilingual extraction of comparable documents

Manuela Yapomo[1], Gloria Corpas[2], Ruslan Mitkov[3]
[1]Evaluations and Language resources Distribution Agency (ELDA), [2]University of Malaga, [3]University of Wolverhampton
manuyap@yahoo.fr, gcorpas@uma.es, R.Mitkov@wlv.ac.uk

1. Introduction

Comparable Corpora (Skadina et al., 2010a)
a collection of documents gathered according to a set of criteria, e.g. the same proportions of texts of the same genre in the same domain from the same period (McEnery and Xiao, 2007) in more than one language or variety of languages (EAGLES, 1996) that contain overlapping information (Hewavitharana and Vogel, 2008)

Objective:
the development of a tool based on cross-language retrieval which given an input of source collection, outputs a target collection of the 'most comparable' texts to the given source documents. We experiment with English and French.

Rationale:
Comparable corpora have enjoyed an increasing importance in recent years as their exploitation was found to be a productive alternative to parallel corpora in several fields of Natural Language Processing (NLP) and beyond:
❑ Terminology Extraction (Saralegi, San Vicente and Gurrutxaga, 2008)
❑ Machine Translation (Abdul-Rauf and Schwenk, 2009), etc.

Challenges :
❑ what parameters to include in the measurement of documents comparability
❑ how to evaluate their similarity relying on them.

2. Hypothesis

using the CLIR-based approach (Talvensaari et al., 2007) further, we perform ontology based-query expansion thus exploiting the synonymy relation in WordNet with a view to achieving better efficiency in the retrieval procedure

3. Proposed methodology

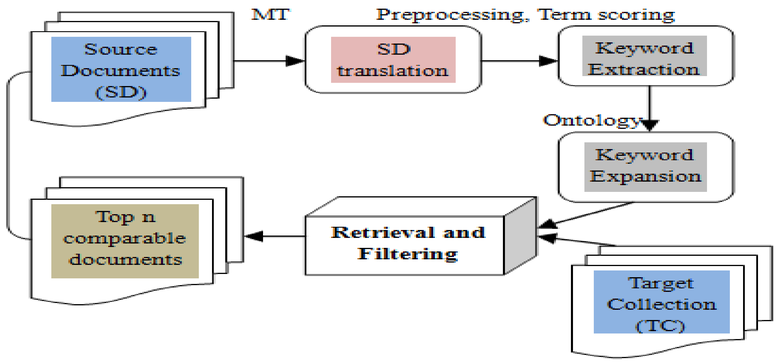

Figure 1: General architecture of the system

The general idea is, given K source documents and M target documents, to extract the N (<=M) target documents most comparable to the source documents.

❑ **Document translation:** The source documents translated into the target language.
❑ **Preprocessing:** Lemmatisation, POS-tagging
❑ **keyword extraction:** tf-idf
❑ **Keyword Expansion:** Synonyms in WordNet, two first lemma-names of the fist synset
❑ **Retrieval and filtering:** topic, time span, text-length

4. Evaluation

Data
❑ Source collection: 38 selected articles in French. with an average number of 659 words per document. domain : "2008 economic crisis
❑ Target set : 280 manually documents annotated according to the following scale

Classes in this study	Equivalent classes according to Braschler and Schäuble (1998)	Comments
Class 1	(1) Same story	The two documents deal with the same event.
Class 2	(2) Related story	Documents deal with the same event or topic from a slightly different viewpoint.
Class 3	(4) Common terminology	Topics are not directly related, but the documents share a considerable amount of terminology.
Class 4	(5) Unrelated	The similarities between the documents are slight/nonexistent.

Table 1: Modification of Braschler and Schäuble 's guidelines for classifying target documents

Results
❑ Retrieval with extracted keywords: performed with average success

	k=10		k=15		k=20	
	#	%	#	%	#	%
Class 1	25	35,7	21	30	18	25,7
Class 2	11	15,7	23	32,8	15	21,4
Class 3	32	45,7	26	37,1	29	41,4
Class 4	2	2,8	0	00	8	11,4
Total	70	100	70	100	70	100

Table 2: Results of retrieval with different sets of relevant keys

❑ Keyword extraction and keyword expansion using WordNet : performance is slightly better only when selecting k2=24

	k1=14		k2=24		k3=31	
	#	%	#	%	#	%
Class 1	20	28,5	21	30	15	21,4
Class 2	13	18,5	24	34,2	12	17,1
Class 3	33	47,1	23	32,8	36	51,1
Class 4	4	5,7	2	2,8	7	10
Total	70	100	70	100	70	100

Table 3: Results of retrieval with different sets of relevant keys and WordNet

K represents extracted kewords in Table 1
K1|2|3 stand for keywords together with associated synonyms

5. Future work

❑ Apply this methodology on much larger sets containing thousands of documents.
❑ Use a more sophisticated term weighting metric for the identification of relevant keys
❑ Exploit more criteria, for example take into account named entities and content descriptors to ensure higher comparability
❑ Use numerical values (a score) to evaluate the degree of similarity of documents

6. Conclusion

❑ This work describes a bilingual approach for extracting comparable documents to a specific set of documents.
❑ Our work takes the CLIR-based approach further. In this study we perform ontology-based query expansion of the most relevant terms thus exploiting the synonymy relation in WordNet
❑ The evaluation of the tool that we developed shows that the best results obtained are after expanding a set to 24 keywords. There is a slight improvement with keyword expansion

7. References

Abdul-Rauf, S. and Schwenk, H. (2009). On the use of Comparable Corpora to improve SMT performance. *Proceedings of the 12th Conference of the European Chapter of the ACL*, Athens , pp.16–23.

Saralegi, X., San Vicente, I. and Gurrutxaga, A. (2008). Automatic extraction of bilingual terms from comparable corpora in a popular science domain. *Proceedings of the Workshop on Comparable Corpora, LREC'08*, Basque Country, pp.27-32.

Talvensaari et al. (2007). Creating and exploiting a comparable corpus in Cross-Language Information Retrieval. *ACM Transactions on Information Systems, 25(1).*

Skadina, I. et al. (2010a). *Analysis and evaluation of comparable corpora for under resourced areas of Machine Translation. Proceedings of the 3rd Workshop on Building and Using Comparable Corpora.* European Language Resources Association (ELRA), La Valletta, Malta, pp.6-1.