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

Abstract

The exploitation of comparable corpora has proven to be a valuable alternative to rare parallel corpora in various Natural Language Processing tasks. Therefore many researchers have stressed the need for large quantities of such corpora and the scarcity of works on their compilation. This paper describes a CLIR-based method for automatic extraction of French-English comparable documents. At the start of the process, source documents are translated and most representative terms are extracted. The resulting keyword list is further enlarged with synonyms on the assumption that keyword expansion might improve the retrieval of such documents. Retrieval is performed on the indexed target collection and a further filtering step based mainly on temporal information and document length takes place. Preliminary results suggest that the employment of ontology could improve the performance of the system.

Keywords: Comparable documents, comparable corpora; Cross-Language Information Retrieval (CLIR); ontology; similarity measurement.

1. Introduction and Previous Work

Comparable corpora are referred to as collections of documents in the same or in different languages made up of similar texts. Using snippets of several definitions, Skadina, et al. (2010a, p.7) came up with a more elaborate description which is the following: "a collection of similar documents that are collected 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 (Munteanu and Marcu, 2005; Hewavitharana and Vogel, 2008)".

The present work focusing on the collection of comparable documents discusses 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. This tool is cross-lingual in its nature as the source and target collections can be in two different languages. In this particular project, we have experimented with English and French.

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. Several works on terminology extraction (Gamallo, 2007; Saralegi, San Vicente and Gurrutxaga, 2008), Machine Translation (MT) (Munteanu and Marcu, 2005; Abdul-Rauf and Schwenk, 2009), Cross-Language Information Retrieval (CLIR) (Talvensaari et al., 2007), etc. relying on comparable corpora provide empirical evidence for this view. They play an important role for translation and terminology as well (Bowker and Corpas, forthcoming).

Comparable documents are traditionally acquired from the web or from existing research corpora and different approaches have been proposed to perform this task. To mine English-German-Spanish comparable documents from the Internet, Talvensaari et al. (2008) employ focused crawling. Domain specific vocabulary is collected separately in all three languages and used to acquire relevant seed URLs. The selected URLs are then employed in the crawling phase to identify relevant pages from which text paragraphs are extracted. Leturia, San Vicente and Saralegi (2009) present a search engine-based approach for acquiring specialised Basque-English comparable corpora from the web. The tool takes as input a mini-corpus from which most relevant words are extracted and used as seeds to retrieve relevant web pages. Relying on two newspaper subcorpora, Bekavac et al. (2004) describe the collection of Bulgarian-Croatian comparable documents by mapping common vocabulary and publication dates in documents of the two corpora. Talvensaari et al. (2007) introduce the CLIR-based approach in gathering comparable Swedish-English documents from two newspaper collections. They extract good keys with RAFT (Relative Average Term Frequency). The resulting keys are translated and ran against the target collection with Lemur retrieval system (www.lemurproject.org).

Our work takes the CLIR-based approach further. In this study, we perform ontology based-query expansion thus exploiting the synonymy relation in WordNet with a view to achieving better efficiency in the retrieval procedure. This novel approach is applied to the bilingual compilation of comparable documents in English and French. The general idea of our methodology is, given K source documents and M target documents, to extract the N (<=M) target documents most comparable to the source documents. Applying this methodology in an incremental fashion would be the basis of compiling comparable corpora.

The paper is organised as follows: Section 2 describes our methodology and outlines the system architecture. Section 3 reports the evaluation results obtained so far with regard to the performance of the system. Finally, section 4 offers concluding remarks.

2. Methodology and Architecture of the System

The source documents are first translated into the target language. They then undergo preprocessing prior to keyword extraction. The list of keywords obtained is further expanded with synonyms. After the phases of document translation, keyword extraction and expansion, document retrieval and filtering are undertaken. The pipeline of the system is illustrated in Figure 1:

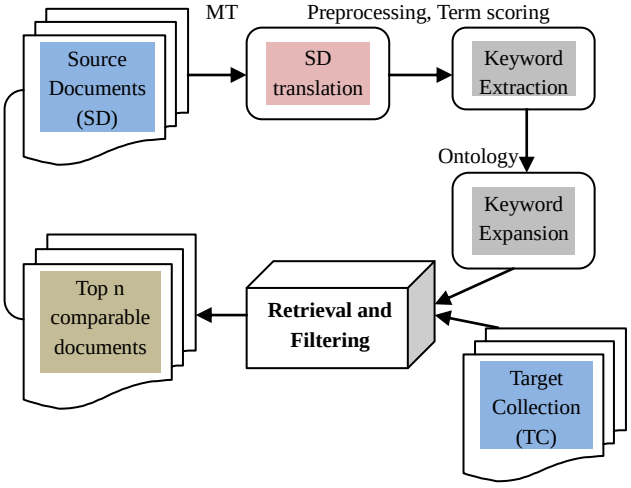

Figure 1: General architecture of the system

2.1 Document Translation

Cross-language retrieval research so far has exploited either dictionary translation (Pirkola et al., 2001) or Machine Translation (Huang et al., 2010). Each translation approach has its advantages and disadvantages. For queries -which are list of words,- dictionary translation appears to be more appropriate. In multilingual dictionaries however, words carry usually more than one translation, and thus ambiguity is carried over to the target language.

In general, MT usually produces a better translation than dictionary-based translation as syntax and other factors are usually taken into account (depending on the MT system). As a result, there is less ambiguity in a translation performed by an MT system. However, the performance of an MT system may not always be of acceptable quality. In general, there is consensus that MT is more suitable for document translation than for keywords translation. However, as in dictionaries, OOV (Out Of Vocabulary) words are encountered with MT tools which also often miss domain-specific terminology. In this work we employ MT based on the premise that it works better for document translation and helps avoiding the problem of ambiguity occurring with dictionaries. Microsoft Translator has been selected as an MT system for this study. The output of the MT system is subject to further processing, namely keywords extraction.

2.2 Keyword Extraction

Prior to performing keywords extraction, the system performs (i) preprocessing of data and (ii) term weighting. Preprocessing in the present study consists in lemmatisation and POS-tagging using the TreeTagger (Schmid, 1994), a tool for annotating texts with part-of-speech and lemma information. Lemmatisation is performed to transform inflected forms into their base forms. POS-tagging is a better alternative to stop words removal as only content words, which are nouns, proper nouns, adjectives and verbs are taken into account. Lemmatisation is a further advantage for languages such as French, which has a rich flexive system. It helps avoiding incorrect count of a term frequency for words which have more than 1 part-of-speech tag. POS-tagging is also helpful in decreasing ambiguity of multi-category words in WordNet.

The next step of term weighting consists in assigning a relevance value to content-bearing words in the source collection. A number of approaches have been proposed to this end. They can be grouped as supervised and unsupervised methods. Supervised methods involve machine learning (Zhang et al., 2006). They are quite stable but demand much effort, since training annotated corpus and a classifier are required. In this work, unsupervised methods are preferred to supervised ones. Following this approach, several formulae have been proposed.

Word frequency or term frequency (TF) was introduced by Luhn (1957) but is quite basic. More robust term weighting methods are preferable. Matsuo and Ishizuka (2004) used word co-occurrence to identify keywords from a unique document. TF-IDF is a standard relevance measure used in several studies (Ramos, 2003; Li, Fan and Zhang, 2007). A limitation of TF-IDF is that it does not necessarily show the goodness of relevant keys that may occur just once or twice in some important documents. Furthermore, the collection should be large enough to yield a reliable IDF. Since our source documents meet the previous requirement for IDF, we will adopt TF-IDF as relevance measure in this work.

After weight is assigned to all the content bearing words in our source documents set, we can move on to keyword extraction. This will be done by selecting the top n keys with higher TF-IDF values. We can proceed to keyword expansion, which we believe might increase the performance of the system.

2.3 Keyword Expansion

Keyword expansion consists in enlarging a keyword list. This is done by adding to the list of initial keywords, words with which they share some semantic relations. Approaches to keyword expansion are based on

probabilistic and ontology-based methods. Probabilistic query expansion consists in extracting terms that are most related to query keys based on co-occurrences of terms in documents. The ontology-based method, on the other hand, makes use of semantic relations already established in ontologies to select terms. In this work, we are interested in this latter approach to keywords expansion. We exploit synonymy in Wordnet (Miller et al., 1993).
How to expand queries automatically is not a trivial task because one has to avoid the problem of ambiguity. When integrating WordNet in our system, we attempt to resolve this problem by POS-tagging our source collection. In this way, the POS-tag could help discarding other categories of a polysemous word. In other to further reduce ambiguity, we will select only the first synset (synonym set) of a word. The choice of the first synset is quite simplistic but will work in most cases for it is the most general sense. We also limit ourselves to the two first lemma-names of the first synset in other to avoid proliferation of keywords.

2.4 Retrieval and Filtering

Document retrieval can be referred to as the matching of some query against a collection of texts with the purpose of obtaining documents relevant to the query only. In line with the definition of comparable corpora in section 1, not only similarity of target documents to the query will be taken into account but also temporal information and size of related documents in our objective to retrieve comparable documents .
In this work, the Opensource toolkit Indri is used to carry out the retrieval process. Indri is part of the Lemur project. Prior to document retrieval, all the target documents were indexed with Lemur. Date normalisation is equally performed according to a specific date format understandable by Indri toolkit. After indexing, proper retrieval can be undertaken. To do filtering based on extralinguistic criteria (date of publication and document length), the corresponding feature-intervals should be defined so as to select only documents that meet the filtering constraints adopted. Since this tool should work with any linguistic data, time span will be extracted from the source documents to ensure that all filtered documents fall within the same time-period and have a text-length ranging from 1,000 to 50,000 characters. This interval is mainly chosen to filter out too small and too large documents.

3. Evaluation

In this part of the paper, we first describe the data that will be used for tests. Experiments and results are then reported with observations.

3.1 Data

To carry out experiments, we use two sets of source and target documents made up of news articles, randomly collected from different news websites.
Our source collection contains 38 selected articles in French. The criteria to meet when selecting the texts are that they should be about the same or closely related topic. The total number of words contained in our source set is of 25,047 with an average number of 659 words in each document. The domain of selected documents was economy and they were all more or less related to the topic of "2008 economic crisis" Documents were taken from news websites lemonde.fr, lepoint.fr, etc.

As regards the target document set we selected 280 which we classified. We opted for a modified version of Braschler and Schäuble (1998)'s relevance scheme as comparability metric for annotation and evaluation purposes. Table 1 illustrates our modification of Braschler and Schäuble's relevance 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	The two documents deal with the same event or topic from a slightly different viewpoint. Alternatively, the other document may concern the same event or topic, but the topic is only a part of a broader story or the article is comprised of multiple stories.
Class 3	(4) Common terminology	The events or 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 or nonexistent.

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

Our modification of Braschler and Schäuble's scheme consists in the deletion of the third class (shared aspects) on the grounds that named entities are not taken into account in our study. Retrieved documents belonging to Class 1 and 2 are considered good alignments whereas retrieval of documents from class 3 and 4 is not.
To classify documents at hand, precisions were added as regards the theme of the documents collection for our experiments:

(1) *Same story* in this context contains texts that are about *the Great Recession*. This includes texts

about causes, manifestations and effects; descriptive, explanatory texts, etc.

(2) *Related story* involves documents reporting financial crisis. It includes articles about financial crises in general or specific ones, different from that of the first category. Examples are *the Great Depression* or *Inflation in Zimbabwe.*

(3) *Common terminology* comprises documents sharing vocabulary. These are documents which are about finances in general.

The documents collected were distributed in each class as illustrated in Table 2 below:

Collection	# of documents	Class	Time Span
Source set (Fr)	38	Class 1	2007 – 2011
Target set (En) (280)	69	Class 1	2007 – 2011
	63	Class 2	No date and size restriction
	81	Class 3	
	67	Class 4	

Table 2: Description of source and target data

3.2 Experiments

We evaluated the performance of our tool on the data described in the previous section. To achieve the retrieval of comparable documents, we had to extract keywords from a translation of source documents using TF-IDF. We further exploited WordNet to enlarge the keyword list with synonyms. The resulting translated keys were used as queries and run against the target language data with Lemur retrieval system. Date of publication and size are used to further filter out less relevant documents.

Experiments were carried out with different configurations to find out which one gives the best results. Different options were tried at the levels of (i) keyword extraction and (ii) keyword expansion. Our experiments can be split in two groups. The purpose of our first group of experiments was to determine which portion of most relevant keys (k) was to be used for retrieval. We carried out experiments with k=10, k=15 and k=20 respectively. Keyword extraction performed with average success. Among the extracted keys, good ones perfectly matching the topic were *recession, subprime*. Relatively good keys were *bankruptcy, mortgage, price, lending, bank*. Many irrelevant keys such as *institution, country, recover, down* were extracted which would negatively affect retrieval. Relevant words such as *crisis, economy, deflation*, etc were not extracted.

In the second set of experiments, we tested the effect of WordNet as described in section 2.3. After expansion of keywords lists k=10, k=15 and k=20, we respectively obtained the following expanded lists k1=14, k2=24 and k3=31 terms. Most of the words in the initial keyword list did not find synonyms in WordNet and most of those that were assigned synonyms were not good keys. Some are *institution (establishment), country (state, land), recover* (*regain, find*).

In the two different groups of experiments, time span and size are used to further filter out documents. As mentioned in section 2.4, temporal information is extracted from source data if available and a size interval of 1,000 to 50,000 characters of texts always applies.

3.3 Results

To carry out evaluation of the efficiency of the system designed, we analyse results of retrieval carried out in the two sets of experiments described in the previous section.

Table 3 shows results of retrieval using different sets of significant terms.

	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 3: Results of retrieval with different sets of relevant keys

The shaded areas in Table 3 and Table 4 below show the best retrieval performances for classes 1 and 2. Results of retrieval show that most of the documents retrieved belong to class 3. This can be explained by the fact that keys extracted are very general words in the semantic field of finance.

Few documents of the second class were retrieved contrarily to documents of the third class which are less comparable. This may be due to the presence of very general words in the keywords list. Around 30% of retrieved documents fall within class 1. We can observe than the first and second sets of keywords, k=10 and k=15 perform better for retrieval of class 1 documents. The second set of keys (k=15) allows retrieval of the largest amount of documents in class 2.

Table 4 shows results of retrieval with the same set of words as those in Table 3 with the difference that keywords are now expanded with synonyms in WordNet.

	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 4: Results of retrieval with different sets of relevant keys and WordNet

With keyword expansion, retrieval appears to be less efficient for documents of class 1. Similarly to the previous group of experiments, more documents from the third class are extracted. The experiment with k2 performs best. Indeed, with this scheme, fewer documents from the third class are extracted and more documents from the second class are obtained.

Though we cannot formulate general conclusions based on these results from our small set of data, we observe that the best results were obtained using the top 15 keys with synonyms in WordNet. WordNet therefore seems to have a positive impact on the retrieval.

4. Conclusion

This work describes a bilingual approach for extracting comparable documents to a specific set of documents. Given K source documents, the N (<=M) most comparable documents to the source documents are extracted from an M target set. Applying this methodology in an incremental fashion would be the basis of compiling comparable corpora.

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 with a view to achieving better efficiency in the retrieval procedure. The evaluation of the tool that we developed shows that the best results obtained are after expanding a set to 24 keywords.

5. 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.

Bekavac, B., Osenova, P., Simov, K. and Tadić, M. (2004). Making monolingual corpora comparable: a case study of Bulgarian and Croatian. *Proceedings of the 4th Language Resources and Evaluation Conference: LREC04*, Lisbon, pp. 1187-1190.

Bowker, L. and Corpas, G. Translation Technology. In Mitkov, R. *The Oxford Handbook of Computational Linguistics*. Second, substantially revised edition. Oxford Unversity Press,

Braschler, M. and Schäuble, P. (1998). Multilingual information retrieval based on document alignment techniques. *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries*. Berlin: Springer- Verlag, pp.183–197.

Gamallo, P. (2007). Learning bilingual lexicons from comparable English and Spanish corpora. *Proceedings of Machine Translation Summit XI*, Copenhagen, pp. 191-198.

Huang, D., Zhao, L., Li, L. and Yu, H. (2010). Mining large-scale comparable corpora from Chinese-English news collections. *Proceedings of the 22th International Conference on Computational Linguistics:* Coling 2010, Beijing, August 2010, pp. 472–480;

Leturia, I., San Vicente, I. and Saralegi, X. (2009). Search engine based approaches for collecting domain-specific Basque-English comparable corpora from the internet. *5th International Web as Corpus (WAC5)*. Donostia-San Sebastian.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D. and Miller K. (1993). Introduction to WordNet: An on-line lexical database. Cambridge: MIT Press.

Munteanu, D. and Marcu, D. (2005). Improving Machine Translation performance by Exploiting non-parallel corpora. *Journal Computational Linguistics, 31(4)*. Cambridge: MIT Press, pp.477-504.

Pirkola, A., Hedlund, T., Keskustalo, H. and Järvelin, K. (2001). Dictionary-based Cross-Language Information Retrieval: Problems, Methods, and Research Findings. *Information Retrieval, 4(3-4)*, pp.209-230.

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.

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.

Schmid, H. (1994). Part-of-Speech tagging with Neural Networks. *Proceedings of the 15th International Conference on Computational Linguistics: COLING-94*.

Talvensaari, T., Laurikkala, J., Järvelin, K., Juhola, M. and Keskustalo, H. (2008). Focused web crawling in the acquisition of comparable corpora. *Information Retrieval 11*, pp.427-445.

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

[1] **www.lemurproject.org** (Accessed February 17, 2012).