Automatic Compilation of Comparable corpora

Manuela Yapomo

Automatic Compilation of Comparable Corpora

Manuela YAPOMO
Centre Tesnière, Université de Franche-Comté, France
Research Group in Computational Linguistics, University of Wolverhampton, United Kingdom

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. Our purpose in this paper is to address this issue by using the CLIR-based method for the automatic acquisition 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. Results are fair and suggest that the use of ontology may improve the performance of the system.

Key-words

Comparable corpora; Cross-language information retrieval; (non-)linguistic criteria; similarity measurement.

1. Introduction and Previous Work

Comparable corpora are referred to as collections of documents in different languages or language varieties made up of similar texts. The present work is about conceiving a program for the automatic cross-language retrieval, within a target collection, of texts most comparable to given source documents. This activity aims at the acquisition of a bilingual comparable corpus.

Comparable corpora have enjoyed an increasing importance in previous years as their exploitation was found to be a productive alternative to parallel corpora in several fields of Natural Language Processing (NLP). 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.

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 used in the crawling phase to identify relevant pages from which text paragraphs are extracted. Leturia, San Vicente and Saralegi (2009) rather 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, Osenova, Simov and Tadić (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 is a further enterprise using the CLIR-based approach. We are interested in the acquisition of French-English comparable corpora.

The paper is organized in the following way: Section 2 describes our methodology by depicting each component of the system architecture. Section 3 is an evaluation of the performance of the system conceived, followed bythe conclusion in section 4.

2. Architecture of the System

The starting point of our methodology is a set of source documents. This linguistic data is first translated into the target language. They then undergo preprocessing prior to keyword extraction. The list of keywords obtained is further enlarged with synonyms. After the phases of document translation, keyword extraction and expansion, document retrieval and filtering can be undertaken. The process is illustrated in Figure 1:

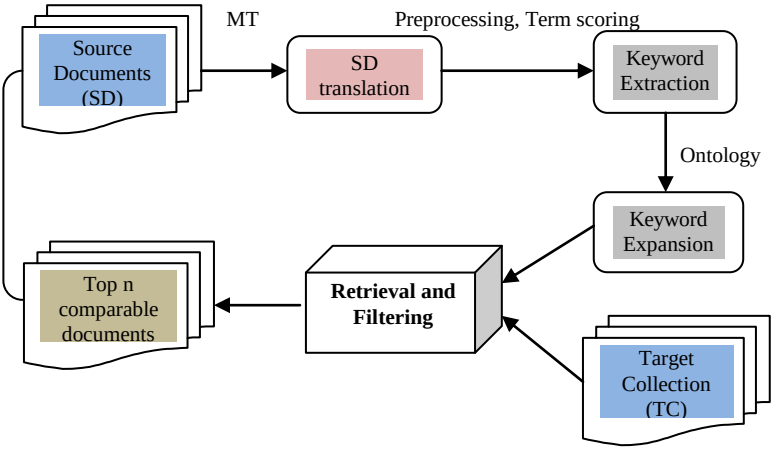

Figure 1. General architecture of the system

2.1 Document Translation

In this work focusing on cross-language retrieval of comparable documents, translation is a key stage. There are general approaches used in automatic translation. These are dictionary translation, machine translation, parallel corpora and comparable corpora translations. The two first methods have been used in various works (Pirkola, Hedlund, Keskustalo and Järvelin, 2001; Huang, Zhao, Li and Yu, 2010).
They both present some advantages and disadvantages. For queries – which are list of words -, dictionary translation seems more appropriate. In

multiilingual dictionaries, a word can be ambiguous in the target language and thus bear several translations. The dictionary method therefore poses the problem of ambiguity.
MT usually produces the best translation in that it takes into account more parameters than dictionaries such as context to determine the most suitable translation of a word. This sharply decreases translation ambiguity which is a considerable problem with dictionaries. However, the single translation returned by an MT system might not be the good one. MT is more suitable for document translation than for keywords translation but as dictionaries, OOV (Out Of Vocabulary) words are encountered and they often miss domain-specific terminology.
We use MT in this work for it works better for document translation and helps avoiding the problem of ambiguity occurring with dictionaries. Microsoft Translator will be used at this level and the resulting translation will be the input for further processing, namely keywords extraction.

2.2 Keyword Extraction

Prior to performing keywords extraction, two tasks are to be undertaken. These are (1) preprocessing of data and (2) term weighting.
Preporcessing 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 will be useful as an alternative to stop words removal. Only content words, which are nouns, proper nouns, adjectives and verbs will be taken into account. Another advantage of this preprocessing stage is that it helps avoiding wrong count of a term frequency for multi-category words. POS-tagging will equally be very useful as a way to decrease 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, Xu, Tang and Li, 2006). They are quite stable but demand much effort, since a training annotated corpus and a classifier are required. In this work, unsupervised methods are preferred to supervised ones. Following this approach, several formulas have been conceived.

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 limit 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 big enough for 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 mentioned in the literature are the 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 rather 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 and may not always be appropriate 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 our comparable documents retrieval, not only similarity of target documents to the query will be taken into account but also temporal information and size of related documents.
In this work, Opensource toolkit Indri is introduced to carry out the retrieval process. Indri is part of the Lemur project. On the basis of Lemur, it combines inference networks with language modeling. Prior to document retrieval, all the target documents were indexed with Lemur. Date normalisation is equally performed for the Indri toolkit understands specific date formats. After indexing, the central task of retrieval could be performed. In filtering based on extralinguistic criteria– date of publication and document length -, intervals will be defined in other to select only documents which conform to them. Since this tool should work with any linguistic data, time span will be extracted from the source document to ensure that documents fall within the same period and a length interval of 1,000 to 50,000 charaters always applies.

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 gathered from different news, government websites, etc.

Our source collection contains 38 manually collected articles in French. The criteria to meet when collecting 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.
As regards the target document set, it is composed of manually collected and classified 280 documents. To assess the performance of our tool in retrieving relevant comparable documents, we had to collect and classify the documents in a particular way. The relevance scale used in collecting, annotating and evaluating the quality of retrieved documents and hence the efficiency of our methodology is a slight modification of Braschler and Schäuble (1998). Table 1 illustrates the relevance scale used in this work:

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. Guidelines for classifying target documents

In Table 2, our modification of Braschler and Schäuble (1998) consists in the deletion of the third class (shared aspects) on the grounds that named entities will not be taken into account in our study. Retrieved documents belonging to Class 1 and 2 are considered good alignments whether 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

In this section of our work, we are to assess the efficiency of our tool with 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 alternatives to find out which one gives the best results. Different options were tried at the levels of (1) keyword extraction and (2) keyword expansion. Our experiments can be split into two parts. 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 fairly. 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 below shows the 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	7	100

Table 3. Results of retrieval with different sets of relevant keys

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 about finances.

Few documents of the second class were retrieved contrarily to documents of the third class which are less comparable. This may be due the presence of very general words in the keywords list. With regards to the number of documents retrieved belonging to the first class, the second and third sets of keys perform better. 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.

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	20	100	20	100	20	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 previous 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

The present work, triggered by the lack of effective methods for the compilation of comparable corpora which are vital in several fields of NLP and linguistics, aimed at building a tool working in a bilingual mode, for the automatic collection of comparable documents. To achieve this, we used the CLIR-based approach.

In order to automatically identify keys relevant to the theme, in our source set, terms were weighted with TF-IDF. We extracted most relevant keys which were used with synonyms in retrieval. We then filter out documents that do not match the time span defined by the source set and a specific length interval. From our experiments using initial keywords, we obtained relatively tolerable retrieval results. We noticed that the use of the two smaller sets of words – 10 and 15 – slightly outperformed the use of 20 keys. The use of WordNet as proposed in this work suggested that it could increase the performance of the system but more tests are to be performed.

Some limitations can be identified in this study. First, the fact that relatively small data was used in the experiments questions the reliability of results and hence conclusions drawn from them. It will be desirable to apply this methodology on much larger sets containing thousands of documents. As regards the acquisition of keywords, our use of tf-idf and WordNet for keywords extraction and expansion respectively is quite simplistic. Using a more sophisticated term weighting metric and improving our exploitation of the WordNet ontology are key future tasks.

In the future, we will equally look forward to exploiting more criteria. We could for example take into account named entities and n-grams.

Acknowledgements

This project was supported by the European Commission, Education & Training, Eramus Mundus: EMMC 2008-0083, Erasmus Mundus Masters in NLP & HLT programme.

I sincerely thank my supervisor Professor Ruslan Mitkov, from the Research Institute in Information and Language Processing at University of Wolverhampton, UK. I also address special thanks to my co-supervisor Professor Sylviane Cardey of the research centre in Linguistics and Natural Language Processing, Lucien Tesnière, University of Franche, Comté, France.

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.

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.

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.

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

Talvensaari, T. et al., 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 12 May 2011).