Multilingual document clusters discovery

Benoit Mathieu & Romaric Besançon & Christian Fluhr
CEA-LIC2M
B.P.6 92265 Fontenay-aux-roses Cedex France
{mathieub,besanconr,flurhc}@zoe.cea.fr

Abstract

Cross Language Information Retrieval community has brought up search engines over multilingual corpora, and multilingual text categorization systems. In this paper, we focus on the multilingual clusters discovery problem, which aim is to extract topic-related multilingual document clusters from a multilingual document collection in an unsupervised way. Our approach is based on a linguistic analysis of the documents that allows to identify relevant features for a vector representation of the documents, each language being associated with a different vector space. We propose a cross-lingual similarity measure for the documents, using bilingual dictionaries. A Shared Nearest Neighbor clustering algorithm is then used to build the clusters. We present an evaluation framework for this task, analyze and discuss the results we obtained and propose directions for future works.

Résumé

En recherche d'information multilingue, beaucoup de travaux ont été menés sur les moteurs de recherche et sur les systèmes de catégorisation automatique de texte. Dans cet article, nous abordons le problème de découverte automatique de classes de documents dans une collection multilingue. Le but est d'extraire des documents de langues différentes traitant d'un même sujet. Notre approche se fonde sur une analyse linguistique des documents permettant d'obtenir une représentation vectorielle des documents, chaque langue étant représentée dans un espace vectoriel différent. Nous proposons une mesure de similarité cross-lingue sur ces documents, utilisant des dictionnaires bilingues. Un algorithme de classification du type 'plus proches voisins' est alors utilisé pour construire les classes. Nous présentons également une méthode d'évaluation ainsi que les résultats obtenus. Ces résultats seront discutés et nous envisagerons plusieurs axes d'amélioration du système.

1. Introduction

Information technology and globalization generate more and more electronic documents written in different languages. Dealing with such an amount of data requires automatic systems to filter, retrieve and classify multilingual documents. The Cross Language Information Retrieval community has designed multilingual search engines and text categorization systems that help users with specific needs. Search engines retrieve documents related to a specific user need expressed by a query, text categorization systems are used to assign to each document one of several predefined categories.

Several approaches have been considered to solve the cross-language information retrieval problem. Search engines translate the user query in all indexed languages, then retrieve documents of different languages and merge results (Savoy J. 2002; Fluhr C. et al. 1997). Query translation can be achieved by the simple use of dictionaries (no disambiguation), or by machine translation (with disambiguation). An alternative way for query translation is Latent Semantic Indexing (Littman M. et al. 1997), which

represent documents in a language independent semantic space, built from the vector representation of parallel texts.

Multilingual text categorization systems have developed solutions such as translating documents in all languages, either manually or automatically (Jalam R. et al. 2004), or using a thesaurus like Eurovoc (Steinberger et al. 2002), that allows identifying the same concepts over different languages. In the monolingual case, several text categorization methods have been evaluated (Yiming Y. 1999). The Nearest Neighbor approach provides good results and seems to be the more scalable method. Jalam R. (2004) proposed a framework for multilingual text categorization. He also evaluated a straightforward method that consists in automatically translating documents in a pivot language.

In this paper, we focus on the unsupervised discovery of multilingual document clusters in a document collection. A multilingual document cluster is a set of documents, possibly written in different languages, which are related to the same topic. Our aim is to extract important topics that exist in a multilingual document collection.

This task differs from information retrieval in that we do not consider specific queries that define the relevant topics, and we have to compare each document with the others in a symmetric way. This task differs from text categorization because the topics identified by the clustering phase are not predefined, and a document can be assigned to more than one cluster.

In the monolingual case, document clustering has been studied extensively (Steinbach M. et al. 2000; Pantel P., Lin D. 2002). Ertoz L. et al. (2001) proposed a Shared Nearest Neighbor approach to manage this task in a monolingual environment. Applying the same method to a multilingual document collection requires us to use a cross-lingual document similarity measure. In the multilingual case, Silva J. et al. (2001) proposed a method for multilingual document clustering based on Relevant Expressions (RE), extracted from the documents and used as base features for the clustering, but the clusters obtained are monolingual.

Systems for multilingual document summarization also face the problem of document clustering from multilingual document collection. Evans D. and Klavans J. (2003) describe a multilingual version of Columbia Newsblaster, which collects news in different languages from multiples sites, extract topics and summarize. To deal with the language gap, they perform automatic translation of documents into English.

Our main focus in this paper is the definition of a cross-lingual similarity measure for comparing documents in different languages. The clustering system is based on a linguistic analysis of each document in its proper language, that allows to extract language-specific features used for a vector-space representation of the documents (in their proper language space) and a Shared Nearest Neighbor (SNN) algorithm for the clustering, that uses the cross-language similarity measure. The system is presented in more details in section 2. We propose in section 3 an evaluation methodology of the system based on precision and recall, and discuss the results obtained and propose directions for future work.

2. Multilingual cluster discovery

2.1 The multilingual cluster discovery problem

The purpose of the system considered is to find topic-based document clusters in a multilingual document collection. There are no hypotheses about the spread of documents over languages, neither about the number of languages. The aim is not to build clusters for all documents but to extract strongly related clusters of documents. Some documents may not belong to any cluster, others may belong to more than one cluster.

2.2. Overview of the multilingual cluster discovery system

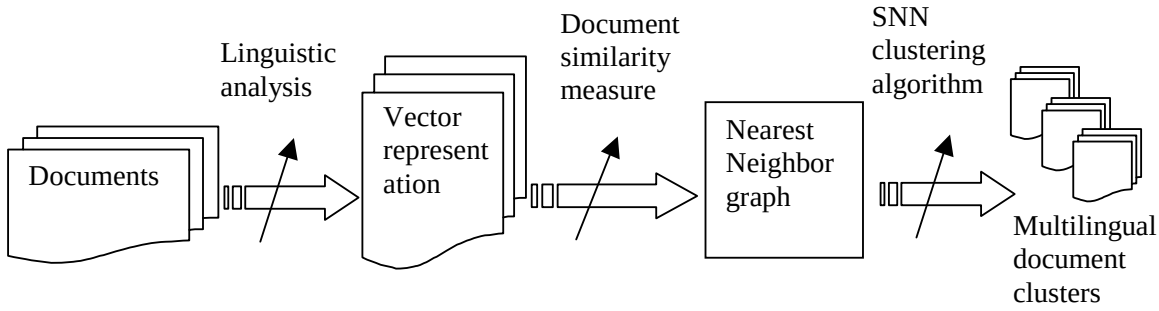

Fig.1. A multilingual cluster discovery system.

We consider that the whole document collection is given to the system. As shown in fig.1, our system is divided in three parts. The linguistic analysis step extracts relevant linguistic units from the documents and builds a vector representation of the documents on the basis of these features. Then, a similarity measure between this vector representation allows the system to build a nearest neighbor graph: the nearest neighbor graph has one node for each document, each node is linked to its *k* nearest neighbors. *k* is the *neighborhood size* parameter. Links are directed, and can be weighted with the similarity between the two documents. Finally, the application of an Shared Nearest Neighbor (SNN) clustering algorithm provide multilingual document clusters.

One can notice that the multilingual behavior of the system highly depends on the similarity measure and particularly its behavior on documents of different languages, but in practice, the SNN clustering algorithm will smooth results and finally provide coherent document clusters.

2.3. Linguistic analysis and document representation

First, the system performs a linguistic analysis of each document (the linguistic processing modules of the system are described in Besançon et al. 2003). After automatic detection of the language of the document, a corresponding linguistic processing is applied (the linguistic processing depends on the language). This processing consists in removing stop words, performing lemmatization, morphosyntactic analysis, and recognizing named entities such as *location*, *organization*, *person*, *time expression*, *numeric expression*, *product* or *event*.

The documents are then represented by a set of terms identified by a triplet (lemma, morphosyntactic category, term type), where the term type can be *keyword* or one of the named entity types. We consider that a term is related to a specific language. The same word in different language is considered as different terms (this often occurs with proper nouns). Later we can assign particular weights for each term type.

We use document frequency to select relevant features among the extracted terms. Typically, we only keep terms that occur in less than 20% of documents of their language. This reduces dimensionality and avoids matching on common terms. Remaining terms are used as base features to represent a document. Each document is then represented by a vector in the vector space corresponding to its own language. The problem is to design a similarity measure that compares the documents represented in different spaces.

2.4. Cosine-like cross-lingual document comparison

As any document collection could contains documents in many several different languages, we did not want to particularize one language and decided to avoid automatic translation of every document in a pivot language (which would allow to represent all documents in a single vector space). We also wanted to propose a solution that does not require having a parallel corpus. We decided to simply use bilingual dictionaries to design a similarity measure of the documents across different languages, so that complete translations of documents in all languages are not required. Dictionaries may contain several translations for polysemous words: in this case, we keep the polysemy and do not try to solve translation ambiguities.

Even if some evaluations of query translation methods in cross-language information retrieval (for instance, Dorr J. and Oard D. 1998) proved dictionary-based translation not to be optimal, the complete translation of all documents in all languages by a machine translation system seem a costly solution. Nevertheless, we plan to evaluate other approaches in the future, such as using a parallel corpus extract statistical information about translation and make the dictionary-based approach more robust. The use of an off-the-shelf machine translation to translate all documents (into one pivot language or into all languages) should also provide a baseline for the evaluation of the similarity measure we propose.

In order to truly obtain multilingual clusters, we need to elaborate a document comparison function that has almost the same behavior in the monolingual and cross lingual cases. We propose here a cosine-like document comparison method (Salton G. and McGill M.J. 1983), with an extension of the TF-IDF weights for the cross-lingual case.

Let us denote:

$T_l = \{t_l\}$: The set of all terms in documents of language l

$oc(t_l, d)$: Occurrences of term t of language l in document d.

$df(t_l)$: Number of documents of language l where t_l occurs.

$nbdoc(l)$: The number of document of language l

Similarity between documents of same language

In this case, d_1 and d_2 are two documents written in the same language l.
The frequency weight of a term t in a document d is defined by

$$tf(t,d) = \begin{cases} 1 + \log(oc(t,d)) & \text{if } oc(t,d) > 0 \\ 0 & \text{otherwise} \end{cases}$$

The inverted document frequency of the term t in language l is defined by

$$idf(t) = \log\left(\frac{nbdoc(l)}{df(t)}\right)$$

The classical cosine similarity function is then defined by

$$sim(d_1,d_2) = \frac{\sum_{t \in d_1 \cap d_2} tf(t,d_1).idf(t).tf(t,d_2).idf(t)}{\sqrt{\sum_{t \in d_1} (tf(t,d_1).idf(t))^2 . \sum_{t \in d_2} (tf(t,d_2).idf(t))^2}}$$

One important thing to notice is that *tf* is the document related part of term weight whereas *idf* is the

language related part of term weight and does not depend on the document.

Similarity between documents of different languages

To build a cross-lingual document comparison function, we have to enhance the language related part of term weight. We decided to use bilingual dictionaries and to consider a couple of translated terms as a unique term in the cross-language space. So we build an inverted document frequency function for translated pairs.

Let us consider d_1 and d_2 two documents written in different languages l_1 and l_2.
We denote D_{12} the bilingual dictionary for languages l_1 and l_2. D_{12} contains a set of term pairs (t_1, t_2), where t_2 is a possible translation in language l_2 of the term t_1 in language l_1. Notice that terms can be polysemous, and each t_1 or t_2 can appear in several pairs.
Let us denote:

- $trans(d_1,d_2)$ The set of couple (t_1,t_2) where t_2 is the translation in language l_2 of term t_1 in language l_1, t_1 occurs in d_1, and t_2 occurs in d_2. We consider t_2 as the translation of t_1 in the following cases:
  - if the lemma pair exists in the dictionary and both have same morphosyntactic category (or same named entity type);
  - if they have the same lemma and are identified as proper noun or named entity.
- $notrans(d_1,d_2)$ The set of terms of d_1 that have no translation in d_2 (notice that this function is not symmetric).

We define the cross-lingual extension of inverted document frequency for a pair (t_1,t_2), where t_2 is a translation of t_1, by:

$$idf(t_1,t_2)=\log\left(\frac{nbdoc(l_1)+nbdoc(l_2)}{df(t_1)+df(t_2)}\right).$$

Then we replace the monolingual *idf* by the cross-lingual *idf* for each translation pair. The cross-lingual similarity measure is then defined by:

$$sim(d_1,d_2)=\frac{\sum_{(t_1,t_2)\in trans(d_1,d_2)} tf(t_1,d_1).idf(t_1,t_2).tf(t_2,d_2).idf(t_1,t_2)}{\sqrt{\left(\sum_{(t_1,t_2)\in trans(d_1,d_2)}(tf(t_1,d_1).idf(t_1,t_2))^2+\sum_{t_1\in notrans(d_1,d_2)}(tf(t_1,d_1).idf(t_1))^2\right)\left(\sum_{(t_1,t_2)\in trans(d_1,d_2)}(tf(t_2,d_2).idf(t_1,t_2))^2+\sum_{t_2\in notrans(d_2,d_1)}(tf(t_2,d_2).idf(t_2))^2\right)}}$$

This function sum contributions of translated terms using the cross-lingual *idf*. In normalization, one must take care of polysemous terms and terms without translations, in order that the function behave the same as the classical cosine. In the denominator, the summation over translated terms ensures that multiple matches of polysemous terms are taken in account. The summation over the *notrans* term set with the monolingual *idf* take the contributions of terms without translations into account.

One can notice that if the bilingual dictionary is symmetric, the comparison function is symmetric too (the *notrans* function is not symmetric, but it is used in the both ways).

Discussion
This comparison function tends to be homogenous over the monolingual and cross-lingual cases, but it highly depends on the translation dictionary coverage and polysemy. In an "ideal" case, the dictionary would map every term of the first language to exactly one term in the second language, so the comparison function would behave exactly the same when comparing document in one language or document in different languages (such ideal case never happens in a real application).

A drawback of this function is that we cannot normalize the document vectors to avoid computing the norm for each comparison because the norm depends on the translations found between the two documents. So, in practice, this function needs much more computation in the cross language case than in the monolingual one.

Using the document comparison function, we can build a nearest neighbor graph with at most $n^2/2$ comparisons (this can be reduced using an index). This step is the main bottleneck of our multilingual cluster discovery system.

2.5. Shared Nearest Neighbor clustering

To create the document clusters using this similarity measure, we used the k-SNN clustering algorithm. A complete description of this algorithm can be found in Ertöz et Al. (2001). We present here the outline of the algorithm in order to specify the parameters used in the evaluation. The algorithm works on a nearest neighbors graph, which is the graph connecting each document (as vertex) with to the k most similar documents. The idea of the algorithm is that the more common neighbors two documents have, the more similar they are. It consists in several steps:

1. Weight each bi-directional links with the number of shared neighbors.
2. Tag the better links as *strong link*. We use the *strong link threshold* parameter to determine the rate of strong links (for example the best 20% links).
3. Compute connectivity of each document. Connectivity is the number of strong links connected to the document.
4. Tag nodes with the highest connectivity as *topic*, and the ones with lowest connectivity as *noise* (use a *topic threshold* and a *noise threshold* parameter to determine the rate of topic and noise nodes.)
5. Build clusters with topics documents (merge into the same cluster the *topic* documents whose distance is less than a *merging distance* parameter number of strong links.)
6. For each non-noise document, if it is strongly linked to a topic document, then add it to the corresponding cluster.

This algorithm is linear in time and in space. The result is a collection of document clusters. In these clusters we can find topic documents that are highly representative of their neighborhood (means that they represent a subject shared by several other documents), and aggregated documents that should deal with the same subject.

3. Evaluation

3.1. Document collection

To evaluate our multilingual cluster discovery system, we need a multilingual document collection with annotated topics. This collection should contain documents written in different languages about the same topics and other "noise" documents that should not share same subjects.

To constitute this collection, we used a subset of the multilingual collection of the CLEF'2003 evaluation campaign for cross-lingual information retrieval systems (Peters C. 2003). All documents come from

newspapers or press agencies. We used the topic-annotated documents written in English, French and Spanish to build clusters. As noise documents, we use documents that have been annotated as non-topic. This means that, during the CLEF evaluation, those documents have been retrieved by some participants but do not belong to the considered topic. This kind of noise is very specific because it has already misled search engines.

Topic No	*141*	*142*	*143*	*144*	*145*	*146*	*147*	*148*	*149*	…	*All*
English	0	1	15	1	12	2	21	2	0	…	396
French	1	4	14	2	3	0	2	0	2	…	410
Spanish	2	8	29	1	7	0	2	3	3	…	384

Tab 1. Distribution of topic documents over languages

Topic No	*141*	*142*	*143*	*144*	*145*	*146*	*147*	*148*	*149*	…	*All*
English	12	11	21	13	7	14	17	15	6	…	886
French	16	6	11	10	31	17	11	8	13	…	892
Spanish	18	12	14	12	10	12	14	13	8	…	783

Tab 2. Distribution of noise documents over languages

Tab 1. and Tab 2. show the distribution of topic and noise documents over languages. There are 60 different topics; some topics are represented by 2 or 3 documents whereas others are represented by about 50 documents. There are 2971 documents in the resulting collection, 975 written in English, 1016 in French and 980 in Spanish.

3.2. Evaluation measures

The main criteria used in evaluation are precision and recall. We consider couple of documents, and denote *C* the set of document pairs that belong to the same cluster in the results obtained by our system, and *T* the set of document pairs that belongs to the same topic. Precision and recall are then defined as follows:

$$precision = \frac{Card(C \cap T)}{Card(C)}$$

$$recall = \frac{Card(C \cap T)}{Card(T)}$$

In evaluation, we aimed at maximizing the F-measure $F_1 = \frac{2.precision.recall}{precision+recall}$. But we also watched at other criteria like *purity* of clusters (rate of documents belonging to the most represented topic in the cluster), and checked if the documents of different languages corresponding to the same topic were clustered together.

3.3. Results

The best results were obtained with the following parameters: *neighborhood size* of 40, *strong link threshold* of 40%, *noise threshold* of 40%, *topic threshold* 20% and a *merging distance* of 2. With these parameters, we reached a *precision* of 0.75 and a *recall* of 0.45. The results are presented in Tab 3.

Cluster	*Best topic*	*Purity*	*English*	*French*	*Spanish*
1	179	0.93	9/9/11	5/4/5	16/15/15
2	199	0.84	4/3/3	19/15/15	15/14/14
3	176	0.80	20/16/19	14/9/9	17/16/16
4	164	0.95	0/0/12	35/33/42	7/7/12
5	162	0.76	3/1/1	6/5/6	24/19/19
6	197	0.83	31/20/21	59/55/61	0/0/34
7	153	0.49	1/0/0	4/2/3	36/18/19
8	157	0.93	50/48/58	1/0/1	9/8/14
9	159	0.92	25/22/24	9/9/10	8/8/10
10	181	0.93	27/25/31	0/0/85	0/0/49
11	168	0.43	13/7/7	14/4/4	17/8/8
12	163	0.96	47/45/46	1/1/6	0/0/6
13	197	0.72	0/0/21	0/0/61	36/26/34
14	180	0.96	17/16/17	16/15/15	12/12/12
15	143	0.79	3/1/15	18/14/14	31/26/29
16	181	0.97	0/0/35	50/49/85	49/47/49

Tab 3. Discovered clusters analysis. X/Y/Z means that the cluster has X documents of the corresponding language, Y of them belong to the best topic, and the best topic has Z documents of this language.

3.4. Discussion

Table 3 shows that out system succeeded in finding relevant multilingual clusters: 14 clusters of 16 are multilingual, 14 of 16 has *purity* higher than 0.72 and 8 of 16 covers more than 90% of the documents of the best represented topic.

For two topics, 181 and 197, the system split documents in different languages into two clusters, (10,16) and (6,13). This can be explained by the fact that the cross-lingual comparison function still gives a higher similarity for documents of same language than documents of different languages. As the neighborhood graph takes into account only the 40 best neighbors, if there are more than 40 documents in same language for a topic, other language documents might be ignored. One can increase the neighborhood size, but it also brings more noise for other clusters.

As we expected, the system ignored topics that were not enough represented. Indeed, the SNN clustering algorithm needs neighbors to identify topics. So the smallest cluster has 27 documents.

4. Conclusion and future work

In this paper, we describe a cross-lingual document similarity measure that can be used for the discovery of multilingual clusters in a multilingual text collection. This similarity measure relies on a vector representation of the documents on the basis of features obtained by a linguistic analysis of the documents (we hence need a specific analysis for each language of the collection). Bilingual dictionaries are used to compute the similarity: hence, the system requires a bilingual dictionary for all language pairs in the multilingual collection, but the designed similarity measure should easily extend to the case where we have to use a pivot language when no direct translation dictionary is available for some language pairs.

We evaluate this cross-lingual similarity measure for multilingual clustering using the SNN clustering algorithm to build document clusters, on a test collection extracted from the CLEF'2003 evaluation

corpus. The results obtained are promising, and we can expect several improvements in the future. Even if the cross-lingual similarity measure is designed to behave the same when comparing documents written in the same language and documents written in different ones, our evaluation shows that it still tends to gather in a cluster documents of same language prior to different language ones (even if the clusters obtained are finally multilingual). It should be interesting to examine how to enhance the clustering algorithm to manage heterogeneous data like document of different languages. We also plan to compare this measure with others, such as thesaurus-based similarity measures (Steinberger et al. 2002).

Another direction for improvement is the enrichment of the features used for the vector representation of the documents. The use of a thesaurus or a pre clustering of terms would allow to use complex concepts instead of terms and should highly reduce dimensionality and increase quality of results.

Our system performance also needs improvement: after the linguistic analysis, the discovery phase took about 8' for 3000 documents, and is quadratic in number of documents because of the neighborhood graph computing. This can be improved with a better use of indexes.

5. References

Besançon R., De Chalendar G., Ferret O., Fluhr C., Mesnard O. and Naets H. (2003). The LIC2M's CLEF 2003 System. *Proc. Of CLEF 2003*, pages 83-92.

Ertöz L., Steinbach M. and Kumar V. (2001). Finding Topics in Collections of Documents: A Shared Nearest Neighbor Approach. *Proc. Of Text Mine'01, Workshop on Text Mining, First SIAM Intenational Conference on Data mining*.

Dorr B. and Oard D. (1998). Evaluating resources for query translation in cross-language information retrieval. *In proc. Of the First International Conference on Language Resources and Evaluation*, pages 759-763.

Evans D., Klavans J. (2003). A Platform for Multilingual News Summarization. *Technical report, Columbia University Department of Computer Science.*

Fluhr C., Schmit C., Ortet Ph., Elkateb F., Gurtner K. (1997), SPIRIT-W3, A distributed crosslingual indexing and retrieval engine, *INET'97, Kuala Lumpur*.

Jalam R., Clech J. and Rakotomalala R. (2004). Cadre pour la catégorisation de textes multilingues. *Proc. of JADT'2004 (7èmes Journées internationales d'Analyse statistique de Données Textuelles)*.

Jarvis R. A. and Patrick E. A. (1973). Clustering Using a Similarity Measure Based on Shared Nearest Neighbors. *IEEE Transactions on Computers*, vol.C22, No. 11.

Littman M., Dumais S., and Landauer T. (1997). Automatic Cross-linguistic Information Retrieval using Latent Semantic Indexing. In Grefenstette G. editor, *Cross Language Information Retrieval*, Kluwer.

Pantel P., Lin D. (2002). Document Clustering with Committees. *In proc. Of SIGIR'02*.

Peters C., (Ed.) (2003). Result of the CLEF 2003 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2003 Workshop, 21-22 August, Trondheim, Norway.

Salton G. and McGill M.J. (1983). *Introduction to Modern Information Retrieval*. Published by McGraw Hill.

Savoy J. (2002). Report on CLEF-2001 Experiments: Effective Combined Query-Translation Approach. *Second Workshop of the Cross-Language Evaluation Forum (CLEF 2001).*

Silva J., Mexia J., Coelho C.A., Lopes G (2001). Multilingual Document Clustering, Topic Extraction and Data Transformation. *In Progress in Artificial Intelligence, volume 2258 of Lecture Notes in Artificial Intelligence*, pages 74-87. Springer-Verlag.

Steinbach M., Karypis G. and Kumar V. (2000). A comparison of document clustering techniques. *KDD Workshop on Text Mining 2000.*

Steinberger Ralf, Bruno Pouliquen & Johan Hagman (2002)**.** Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus Eurovoc. In: A. Gelbukh editor, *Computational Linguistics and Intelligent Text Processing, Third International Conference, CICLing'2002.*

Yiming Y. (1999). An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval*, vol.1:69-90