

Multilingual News Document Clustering: Two Algorithms Based on Cognate Named Entities

Soto Montalvo¹, Raquel Martínez², Arantza Casillas³, Víctor Fresno¹

¹ GAVAB Group, URJC

{soto.montalvo,victor.fresno}@urjc.es

² NLP&IR Group, UNED

raquel@lsi.uned.es

³ Dpt. Electricidad y Electrónica, UPV-EHU

arantza.casillas@ehu.es

Abstract. This paper presents an approach for Multilingual News Document Clustering in comparable corpora. We have implemented two algorithms of heuristic nature that follow the approach. They use as unique evidence for clustering the identification of cognate named entities between both sides of the comparable corpora. In addition, no information about the right number of clusters has to be provided to the algorithms. The applicability of the approach only depends on the possibility of identifying cognate named entities between the languages involved in the corpus. The main difference between the two algorithms consists of whether a monolingual clustering phase is applied at first or not. We have tested both algorithms with a comparable corpus of news written in English and Spanish. The performance of both algorithms is slightly different; the one that does not apply the monolingual phase reaches better results. In any case, the obtained results with both algorithms are encouraging and show that the use of cognate named entities can be enough knowledge for deal with multilingual clustering of news documents.

1 Introduction

Multilingual Document Clustering (MDC) involves dividing a set of n documents, written in different languages, into a specified number k of clusters, so that the documents that are similar to other documents will be in the same cluster. Meanwhile a multilingual cluster is composed of documents written in different languages, a monolingual cluster is composed of documents written in one language.

MDC has many applications. The increasing amount of documents written in different languages that are available electronically leads to develop applications to manage that amount of information for filtering, retrieving, and grouping multilingual documents. MDC tools can make easier tasks such as Cross-Lingual Information Retrieval, the training of parameters in Statistics Based Machine Translation, or the Alignment of parallel and non parallel corpora, among others.

MDC systems have developed different solutions to group related documents. On the one hand, the strategies employed can be classified in two main groups:

the ones which use translation technologies, and the ones that transform the document into a language-independent representation. One of the crucial issues regarding the methods based on document or features translation is the correctness of the proper translation. Bilingual resources usually suggest more than one sense for a source word and it is not a trivial task to select the appropriate one. Although word-sense disambiguation methods can be applied, these are not free of errors. On the other hand, methods based on language-independent representation also have limitations. For instance, those based on thesaurus depend on the thesaurus scope. Numbers or dates identification can be appropriate for some types of clustering and documents; however, for other types it could not be so relevant and even it could be a source of noise.

MDC is normally applied with parallel [12] or comparable corpus ([1], [2], [6], [7], [10], [13]). In the case of the comparable corpora, the documents usually are news articles. Considering the approaches based on translation technology, two different strategies are employed: translate the whole document into an anchor language, or translate only some features of the document. Some authors, for example [7], use machine translation techniques to translate the whole document, while others apply the same techniques to translate selected features [10]. On the other hand, authors like [1] translate some selected features of the document consulting a bilingual dictionary. Some approaches first carry out a monolingual clustering in each language, and then they find relations between the obtained clusters generating the multilingual clusters. Other approaches start with a multilingual clustering to look for relations between the documents of all the involved languages. In [2] the authors select different features to carry out some experiments with both approaches.

The strategies that use language-independent representation try to normalize the content of the documents in a language-neutral way; for example: by mapping text contents to an independent knowledge representation, or by recognizing language independent text features inside the documents. Both approaches can be employed isolated or combined. The first approach involves the use of existing multilingual linguistic resources, such as thesaurus, to create a text representation consisting of a set of thesaurus items. In [13] the authors present an approach based on using the multilingual thesaurus *Eurovoc*. The second approach involves the recognition of independent elements. In [6] is presented an approach that exploits the presence of common words among different languages for solving cross language text categorization. In [5] use as document features the named entities as well as the publication date of the document to carry out the multilingual clustering. In this case the MDC is applied in order to align a comparable corpora to obtain a similarity multilingual thesaurus. However, in [4] the author affirms that the NEs themselves are not suitable to be used as features in document clustering. In [12] the authors present a method based on Relevant Expressions (RE). Others works ([9], [14]) combine recognition of independent text features with mapping text contents to a thesaurus.

This paper presents an approach for MDC in comparable corpora. We have implemented two algorithms, both of heuristic nature, that use as unique evi-

dence for clustering the identification of cognate named entities between both sides of the comparable corpora. None of the revised works use as unique evidence for clustering the identification of cognate named entities between both sides of the comparable corpora. One of the main advantages of this approach is that it does not depend on multilingual resources such as dictionaries, machine translation systems, thesaurus or gazetteers. In addition, no information about the right number of clusters has to be provided to the algorithms. The applicability of the approach only depends on the possibility of identifying cognate named entities between the languages involved in the corpus. It could be particularly appropriate for news corpus, where named entities play an important role. The main difference between the two algorithms consists of whether a monolingual clustering phase is applied or not. This allows to determine when is more appropriate the application of the monolingual and multilingual phases, or even if a monolingual phase is needed.

In Section 2 we present our approach for MDC and the two algorithms. Section 3 describes the corpora, as well as the experiments and the results. Finally, Section 4 summarizes the conclusions and the future work.

2 MDC by Cognate NE Identification

We propose an approach based only on cognate Named Entities (NE) identification. The NE categories that we take into account are: PERSON, ORGANIZATION, LOCATION, and MISCELLANY. Other numerical categories such as DATE, TIME or NUMBER are not considered in this work. We think they are less relevant regarding the content of the document. In addition, they can lead to group documents with few content in common.

The approach has two main phases: cognate NE identification which is common to the two algorithms, and clustering. Both phases are described in detail in the following subsections.

2.1 Cognate NE identification

This phase is shared by the two algorithms. It consists of two steps:

1. Detection and classification of the NEs in each side of the corpus separately. In our case we used a corpus with morphosyntactical annotations and the NEs identified and classified.
2. Identification of cognates between the NEs of both sides of the comparable corpus.

In order to identify the cognates between NEs 4 steps are carried out:

- Obtaining two lists of NEs, one for each language.
- Identification of entity mentions in each list. For instance, “Ernesto Zedillo”, “Zedillo”, “Sr. Zedillo” will be considered as the same entity after this step since they refer to the same person. This step is only applied to entities of

PERSON category. The identification of NE mentions, as well as cognate NE, is based on the use of the Levensthein edit-distance function (LD). This measure is obtained by finding the cheapest way to transform one string into another. Transformations are the one-step operations of insertion, deletion and substitution. The result is an integer value that is normalized by the length of the longest string. In addition, constraints regarding the number of words that the NEs are made up as well as the order of the words are applied.

- Identification of cognates between the NEs of both sides of the comparable corpus. It is also based on the LD. In addition, also constraints regarding the number and the order of the words are applied. First, we tried cognate identification only between NEs of the same category (PERSON with PERSON, ...) or between any category and MISCELLANY (PERSON with MISCELLANY, ...). Next, with the rest of NEs that have not been considered as cognate, a next step is applied without the constraint of being to the same category or MISCELLANY. As result of this step a list of corresponding bilingual cognates is obtained.
- The same procedure carried out for obtaining bilingual cognates is used to obtain two more lists of cognates, one per language, between the NEs of the same language.

2.2 Clustering

The two algorithms proposed for the clustering of multilingual news documents are of heuristic nature. Both, in an iterative way, decide the number of clusters.

Bilingual at the End Algorithm (BEA) . BEA consists of 3 main phases: (1) first monolingual clusters creation, (2) monolingual relocation of documents, and (3) bilingual relocation of documents. This algorithm is based on a previous one described in [8].

1. First monolingual clusters creation. Documents in each language are processed separately. News of the same language that have more cognates in common than a threshold are grouped into the same cluster. In addition, at least one of the cognates have to be of a specific category. In this work we have fixed this category to be PERSON. After this phase all documents are assigned to some cluster. Notice that some cluster could have only a document since this one does not comply with the grouping conditions. After this phase two sets of clusters are obtained, one per language. The number of clusters obtained in this phase will be the top limit; the next phases could reduce it.
2. Monolingual relocation of documents. In this phase the documents in each language are processed separately as well. Each document is located in the cluster that contains the most similar document regarding the number of cognates in common, but only if that number is greater than a threshold.

No constraint regarding the NE category is applied. This is an iterative process until no document is relocated. As result of this phase, the number of clusters in each set could be reduced because of the relocation.

3. Bilingual relocation of documents. Finally, both sets of monolingual clusters are merged into one. This process is not carried out by the union of the whole clusters, but by the relocation of documents. The process is similar to the previous one, but with the documents and clusters of both languages.

Bilingual Algorithm (BA) . BA consists of 2 main phases: (1) first bilingual clusters creation, and (2) bilingual relocation of documents.

1. First bilingual clusters creation. This phase is similar to the first phase of BEA but comparing news documents of different languages. After this phase only one set of clusters is obtained.
2. Bilingual relocation of documents. This phase is similar to the third phase of BEA. Therefore, documents are compared among them irrespective of the languages. This is why no later phase is needed.

The thresholds of both algorithms can be customized in order to permit and make the experiments easier. In addition, the parameters customization allows the adaptation to different type of corpus or content. In Section 3.2 the exact values we have used are described.

3 Evaluation

We wanted not only determine whether our approach was successful for MDC or not, but we also wanted to compare if the application of the multilingual comparison only at the end or from the beginning influences the results.

3.1 Corpus

A Comparable Corpus is a collection of similar texts in different languages or in different varieties of a language. In this work we compiled a collection of news written in Spanish and English belonging to the same period of time. The news are categorized and come from the news agency EFE compiled by HERMES project (<http://nlp.uned.es/hermes/index.html>). That collection can be considered like a comparable corpus.

We used two subsets of that collection. In order to test the MDC results we have carried out a manual clustering with each subset. Three persons read every document and grouped them considering the content of each one. The first subset, call *S1*, consists on 63 news, 35 in Spanish and 28 in English. It consists on 8 multilingual and 2 monolingual clusters. The second one, *S2*, is composed of 136 news, 71 in Spanish and 65 in English. It consists on 24 multilingual and 2 monolingual clusters.

In the experimentation process the first subset, *S1*, was used to train the parameters and threshold values; with the second one the best parameters values were applied.

3.2 Experiments and Results with MDC by Cognate NE

The quality of the results is determined by means of an external evaluation measure, the F-measure [11]. This measure compares the human solution with the system one. The F-measure combines the precision and recall measures:

$$F(i, j) = \frac{2 \times \text{Recall}(i, j) \times \text{Precision}(i, j)}{(\text{Precision}(i, j) + \text{Recall}(i, j))}, \quad (1)$$

where $\text{Recall}(i, j) = \frac{n_{ij}}{n_i}$, $\text{Precision}(i, j) = \frac{n_{ij}}{n_j}$, n_{ij} is the number of members of cluster human solution i in cluster j , n_j is the number of members of cluster j and n_i is the number of members of cluster human solution i . For all the clusters:

$$F = \sum_i \frac{n_i}{n} \max\{F(i)\} \quad (2)$$

The closer to 1 the F-measure value the better MCD performance.

The threshold for the LD in order to determine whether two NEs are cognate or not is 0.2, except for entities of ORGANIZATION and LOCATION categories which is 0.3 when they have more than one word. In the first clusters creation phase of both BEA and BA algorithms, one of the constraint refers to the category of at least one of the cognates in common. We realized that this constraint mainly influences in the number of clusters obtained in this phase. However, it has little impact in the resulting clustering after the relocation phases. Therefore, we have fix this category to be PERSON in this experiments. Regarding the thresholds of the phases of both algorithms, after training the thresholds with the collection $S1$ we concluded:

- In BEA algorithm two thresholds are needed: one for the first phase (TH1) and the other for the second and third phases (TH2). The second threshold has more impact in the result than the first one. In fact, with a low value for TH2 (2) the best results are obtained. It seems that using a TH1 relatively high (7, 8, 9) leads to a good first grouping that makes second and third phases more effective. However with lower values for TH1 good f-measure results are obtained as well.
- BA algorithm also needs two thresholds, one per phase. It performs the best clustering with both high and low values for TH1 but with low or medium values for TH2. It seems to be more independent of the threshold values.

Table 1 shows the 10 best results of the application of BEA and BA algorithms to subset $S2$. We run the algorithms with the best parameter set obtained of the experimentation with $S1$. This set was the best set for $S2$ collection as well. The fifth column represents: the number of multilingual clusters of the algorithm result, the number of clusters calculated, and the number of clusters of the human solution. Although none of the results got the exact number of clusters, it is remarkable that the resulting values are close to the right ones.

Step Thresholds			Results	Clusters
Alg.	TH1	TH2	F-measure	Mult./Calcul./Total
BEA	7	2	0.8796	23/30/26
	8	2	0.8708	19/31/26
	9	2	0.8708	19/31/26
	10	2	0.8708	19/31/26
	4	2	0.8600	18/29/26
	5	2	0.8600	18/29/26
	6	2	0.8600	18/29/26
	7	4	0.8594	17/42/26
	3	2	0.8569	18/28/26
	8	4	0.8506	17/43/26
BA	2	3	0.8831	22/33/26
	2	2	0.8831	22/33/26
	2	1	0.8831	22/32/26
	2	0	0.8831	22/32/26
	8	3	0.8770	24/36/26
	8	2	0.8770	24/36/26
	8	1	0.8770	24/35/26
	8	0	0.8770	24/35/26
	2	4	0.8750	22/36/26
	2	5	0.8750	22/36/26

Table 1. MDC results with the BEA and BA Algorithms for cognate NE approach and $S2$ subset

4 Conclusions and Future Work

We have described a novel approach for Multilingual Document Clustering based only on cognate named entities identification. One of the main advantages of this approach is that it does not depend on multilingual resources such as dictionaries, machine translation systems, thesaurus or gazetteers. The only requirement to fulfill is that the languages involved in the corpus have to allow the possibility of identifying cognate named entities. Another advantage of the approach is that it does not need any information about the right number of clusters. In fact, the algorithm calculates it according with the threshold values of the algorithm.

We propose two algorithms that follow our approach. The main difference between them is whether a previous monolingual clustering phase is applied or not. We have tested the two algorithms with a comparable corpus of news written in English and Spanish, obtaining encouraging results. The one that does not apply a monolingual phase obtains slightly better clustering results. This approach could be particularly appropriate for news articles corpus, where named entities play an important role. Even more, when there is no previous evidence of the right number of clusters. Future work will include the compilation of more corpora, the incorporation of machine learning techniques in order to obtain the thresholds more appropriate for different type of corpus.

Acknowledgements

We wish to thank the anonymous reviewers for their helpful and instructive comments. This work has been partially supported by MCyT TIN2005-08943-C02-02.

References

1. B. Mathieu, R. Besancon and C. Fluhr: "Multilingual Document clusters discovery". *RIAO'2004* (2004) 1–10.
2. H-H. Chen and C-J. Lin: "A Multilingual News Summarizer". *Proceedings of 18th International Conference on Computational Linguistics*, (2000), 159–165.
3. G. Karypis: "CLUTO: A Clustering Toolkit". *Technical Report: 02-017*. University of Minnesota, Department of Computer Science, Minneapolis, MN 55455 (2002).
4. W. Gang: "Named Entity Recognition and An Apply on Document Clustering". *MCS Sc Thesis*. Dalhousie University, Faculty of Computer Science, Canada (2004).
5. M. García, F. Martínez, L.A. Urea y M.T. Martín: "Generación de un tesoro multilingüe a partir de un corpus comparable aplicado a CLIR". *Procesamiento de Lenguaje Natural*, vol(28), (2002) 55–62.
6. A. Gliozzo and C. Strapparava: "Cross language Text Categorization by acquiring Multilingual Domain Models from Comparable Corpora". *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, (2005), 9–16.
7. L.J. Leftin: "Newsblaster Russian-English Clustering Performance Analysis". *Columbia computer science Technical Reports* (2003).
8. S. Montalvo, R. Martínez, A. Casillas and V. Fresno: "Multilingual Document Clustering: an Heuristic Approach Based on Cognate Named Entities". To be published in *COLING-ACL 2006* (2006).
9. B. Pouliquen, R. Steinberger, C. Ignat, E. Ksper and I. Temikova: "Multilingual and cross-lingual news topic tracking". *Proceedings of the 20th International Conference on Computational Linguistics*, (2004), 23–27.
10. A. Rauber, M. Dittenbach and D. Merkl: "Towards Automatic Content-Based Organization of Multilingual Digital Libraries: An English, French, and German View of the Russian Information Agency Novosti News". *Third All-Russian Conference Digital Libraries: Advanced Methods and Technologies*, Digital Collections Petrozavodsk, RCDI'2001, (2001).
11. C.J. van Rijsbergen: "Foundations of evaluation". *Journal of Documentation*, vol(30), (1974), 365–373.
12. J. Silva, J. Mexia, C. Coelho and G. Lopes: "A Statistical Approach for Multilingual Document Clustering and Topic Extraction form Clusters". *Pliska Studia Mathematica Bulgarica*, vol(16), (2004), 207–228.
13. R. Steinberger, B. Pouliquen and J. Scheer: "Cross-Lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC". *CICling'2002*, (2002), 415–424.
14. R. Steinberger, B. Pouliquen and C. Ignat: "Exploting multilingual nomenclatures and language-independent text features as an interlingua for cross-lingual text analysis applications". *SILTC* (2004).