Bilingual News Clustering Using Named Entities and Fuzzy Similarity

Soto Montalvo[1], Raquel Martínez[2], Arantza Casillas[3], Víctor Fresno[2]

[1] GAVAB Group, URJC
soto.montalvo@urjc.es
[2] NLP&IR Group, UNED
{raquel,vfresno}@lsi.uned.es
[3] Dpt. Electricidad y Electrónica, UPV-EHU
arantza.casillas@ehu.es

Abstract. This paper is focused on discovering bilingual news clusters in a comparable corpus. Particularly, we deal with the news representation and with the calculation of the similarity between documents. We use as representative features of the news the cognate named entities they contain. One of our main goals consists of proving whether the use of only named entities is a good source of knowledge for multilingual news clustering. In the vectorial news representation we take into account the category of the named entities. In order to determine the similarity between two documents, we propose a new approach based on a fuzzy system, with a knowledge base that tries to incorporate the human knowledge about the importance of the named entities category in the news. We have compared our approach with a traditional one obtaining better results in a comparable corpus with news in Spanish and English.

1 Introduction

The huge amount of documents written in different languages that are available electronically, leads to develop tools to manage such amount of information for filtering, retrieving and grouping purposes. Grouping or clustering automatically related multilingual documents can be a useful task for the processing and management of the multilingual information.

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 most 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 in task 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 is normally applied with parallel [19] or comparable corpus [14], [4], [8], [13], [18], [21]. In the case of the comparable corpora, the documents usually are news articles.

MDC systems have developed different solutions to group related documents. The strategies employed can be classified in two main groups: (1) the ones which use translation technologies, and (2) the ones that transform the document into a language-independent representation.

Regarding the first strategy, some authors use machine translation systems ([13], [2]); whereas others translate the document word by word consulting a bilingual dictionary ([14], [18], [4]). One of the crucial issues regarding the methods based on document or feature 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.

The strategy that uses language-independent representation tries to normalize or standardize the document contents in a language-neutral way; for example: by mapping text contents to an independent knowledge representation, such as thesaurus ([21], [20], [17]), or by recognizing language-independent text features inside the documents ([19], [7], [3]). Both strategies can be either used isolated or combined. 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 of documents or clustering it could not be so relevant and even it could be a source of noise. None of the revised works use as unique knowledge for clustering the cognate named entities shared between both sides of the comparable corpora.

In this work we study two of the crucial issues of MDC: document representation and document similarity calculation. We represent the documents only by means of cognate Named Entities (NE), considering them like independent language features. We propose a document representation by means of different vectors, one per each NE category taken into account. Regarding the similarity, we propose a similarity measure based on a fuzzy rule system, that combines information about the shared NE category. We have tested this approach with a comparable corpus of news written in English and Spanish, obtaining better results than with a traditional approach. We also wanted to confirm preliminary results indicating that the use of only NE is a good source of knowledge for multilingual news clustering since some authors ([5], [6]) suggest that it is not.

In the following section we present our approach for MDC for both, document representation and document similarity calculation phases. Section 3 describes the corpora, as well as the clustering algorithm used, the experiments and results. Finally, Section 4 summarizes the conclusions and the future work.

2 Multilingual News Clustering: Our Approach

Our approach is made up of the combination of two proposals for two outstanding aspects in MDC: document representation and document similarity calculation. In this work, we did not study the clustering algorithms, but we used a partitional one instead. In a previous work [15] we obtained preliminary encouraging MDC results by using only the cognate named entities as news features; therefore,

in this one we have gone more deeply into the study of the representation of news by means of NE. In that work we proposed a clustering algorithm that grouped news according to the number and category of the shared NE among news. That algorithm is highly dependent on the thresholds and the corpora. However, in this work we tried to propose a more robust approach for document representation and comparison, that could be used with any clustering algorithm. The two proposals are described in detail in the following subsections.

2.1 Document representation: Cognate Named Entities Selection

It is well known NE play an important role in news documents. We want to exploit this characteristic by means of considering them like the only distinguishing features of the documents. In addition, we take into account its specific category as well.

We just consider the following NE categories: PERSON, ORGANIZATION and LOCATION. Other categories, such as DATE, TIME or NUMBER are not taken into account in this work because we think they can lead to group documents with few content in common. However, PERSON, ORGANIZATION or LOCATION NE can be suitable to find common content in documents in a multilingual news corpus.

In most of the MDC approaches the document representation is based on the vector-space model. In this model each document is considered to be a vector, where each component represents the weight of a feature in the document. Usually each document is represented with only one vector that contains all the features. Nevertheless, we have generated several feature vectors for each document. In fact, three different partial feature vectors represent each document, one per NE category taken into account: one vector represents the PERSON NE, other vector the LOCATION NE, and the third one the ORGANIZATION NE. Other work where more than one feature vector per document is used is [5]. They use two vectors, one representing the NE and the other one the nouns.

Our proposal allows to use information about the category of the cognate NE that the news can have in common in order to determine the documents similarity. To represent the comparable corpus we only consider the NE that appear in all the languages involved. This decision considerably reduces the number of features taken into account. The cognate NE identification between languages, as well as the PERSON NE coreference resolution is based on the use of the Levenshtein edit-Distance function (LD), described in detail in [15].

Once the vocabulary (set of the cognate NE shared) has been defined, each element of the feature vectors must be weighted using a weighting function. We use the TF-IDF weighting function, which combines *Term Frequency* (TF) and *Inverse Document Frequency* (IDF) to weight terms.

2.2 Document Similarity using a Fuzzy System

Most of the clustering algorithms determine the grouping according to the similarity (or the distance) among the documents. In this case, the similarity

has to be calculated from the three partial feature vectors which represent the documents content by means of the cognate NE.

Other works that represent the content of each document with more than one vector such as [5], fix coefficients for each vector and use different functions in order to carry out linear combinations. In [17] the authors calculate the similarity between clusters in different languages using three vectors with a relative weighted impact of 70%, 20% and 10%, respectively.

In this work, we propose the use of a fuzzy logic system to combine the partial similarities obtained from the three partial vectors. A fuzzy reasoning system sets suitably model the uncertainty inherent to human reasoning processes, by embodying his knowledge and expertise in a set of linguistic expressions that manages words instead of numerical values [11] [9]. Our main aim is to incorporate the human knowledge about the importance of the category of the NE of the news.

Three linguistic variables are defined: PERSON, LOCATION, and ORGANIZATION similarity. Each variable represents the similarity between two documents according to the content of the respective vectors. These partial similarities are considered as the inputs of a fuzzy rule system. The linguistic variable values are inputs in the antecedent of each IF-THEN rule in the knowledge base. The rule consequent has a unique variable: the global DOCUMENT similarity. In Figure 1 we present the input and output of the linguistic variables.
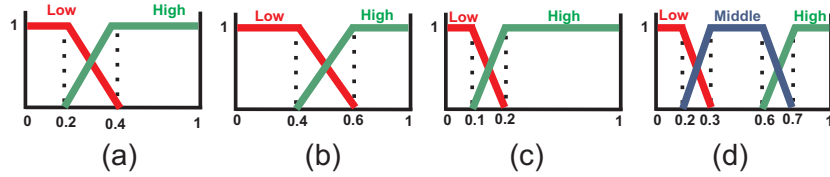

Fig. 1. Linguistic variables: (a) PERSON similarity, (b) LOCATION similarity, (c) ORGANIZATION similarity, (d) DOCUMENT similarity.

The fuzzy set values in the linguistic variables (see values in Figure 1) are defined taken into acco unt the following assumption: we have considered that organization NE are more distinguishing in news, following by person NE, and finally, location NE. For example, is more likely that the same location NE appears in several news of different topic; less probable that a person NE does, and finally, even less that an organization NE does.

The three partial similarities of two documents, corresponding to the input to the three linguistic variables, are obtained with the cosine distance function ($cosine(d_1, d_2)$) normalized to the largest vocabulary dimension: $sim_{part}(d_1, d_2) = cosine(d_1, d_2) \times \frac{dim}{dim_h}$, where dim is the dimension of the partial vectors, and dim_h is the dimension of the largest partial vector.

The knowledge base of the fuzzy system is expressed like a set of IF-THEN rules. These rules try to register the reasoning, or even the common sense that the humans would employ in order to calculate the global similarity between two

news from their partial similarities. In this case, we defined the rules based on the same assumption used to determine the fuzzy set values. With these criteria we defined the following 8 rules:

```
if PS is Low and LS is Low and OS is Low then S is Low
if PS is low and LS is Low and OS is High then S is Middle
if PS is Low and LS is High and OS is Low then S is Middle
if PS is Low and LS is High and OS is High then S is Middle
if PS is High and LS is Low and OS is Low then S is Middle
if PS is High and LS is Low and OS is High then S is Middle
if PS is High and LS is High and OS is Low then S is Middle
if PS is High and LS is High and OS is High then S is High
```

where `PS`, `LS` and `OS` represent the similarity between PERSON NE, LOCATION NE, and ORGANIZATION NE respectively; and `S` represents the resultant document similarity.

3 Evaluation

In this Section, first the corpus is described; next, the clustering algorithm used; and finally, the experiments and the results are presented.

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. The articles belong to a variety of IPTC categories [10], including "politics", "crime law & justice", "disasters & accidents", "sports", "lifestyle & leisure", "social issues", "health", "environmental issues", "science & technology" and "unrest conflicts & war", but without subcategories.

First, we performed a linguistic analysis of each document by means of *Freeling* tool [1]. Specifically we carried out: morpho-syntactic analysis, lemmatization, and recognition and classification of NE; the *NEClassifier* Software [16] is used to detect and classify NEs in the English documents.

We have used five subsets of that collection to evaluate the experiments carried out: $S1$, $S2$, $S3$, $S4$ and $S5$. Subset $S1$ consists of 192 news, 100 in Spanish and 92 in English; subset $S2$ consists of 179 news, 93 in Spanish and 86 in English; subset $S3$ consists of 150 news, 79 in Spanish and 71 in English; subset $S4$ consists of 137 news, 71 in Spanish and 66 in English; and finally, subset $S5$ consists of 63 news, 35 in Spanish and 28 in English. Some articles belong to more than one IPTC category according to the automatic categorization. We were interested in a MDC which goes beyond the high level IPTC categories, making clusters of smaller granularity. So, in order to test the MDC results we carried out a manual clustering with each subset. Three persons read the documents and grouped them considering the content of each one. They judged

independently and only the identical resultant clusters were selected. The human clustering solution of subset $S1$ is composed of 33 multilingual clusters and 2 monolingual clusters; subset $S2$ has 33 multilingual clusters; 26 multilingual clusters has the human solution of subset $S3$; subset $S4$ has 24 multilingual clusters and 2 monolingual; and the solution of subset $S5$ has 8 multilingual clusters and 2 monolingual.

3.2 Clustering Algorithm

Since our objective was not to propose a clustering algorithm, we selected one from the well known CLUTO library [12], the "Direct" algorithm. The input to the "Direct" clustering algorithm is the similarity matrix generated by the fuzzy system, as well as the number of clusters.

3.3 Experiments and Results

The quality of the results are determined by means of an external evaluation measure, the F-measure [22]. This measure compares the human solution with the system one. The F-measure combines the precision and recall measures:

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

where $Recall(i,j) = \frac{n_{ij}}{n_i}$, $Precision(i,j) = \frac{n_{ij}}{n_j}$, 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:

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

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

In order to compare our approach to one traditional, we also represented the documents by means of a vector with all the NE, and we used a well known similarity measure to compute the similarity of two documents: the cosine distance.

The results of the experiments are shown in Table 1. The first column shows the F-measure values for the different subsets of corpus used with the fuzzy approach. The second column shows the F-measure values with the traditional approach, and the third one represent the total number of cognate NE and the number per NE categories used in the corpus representation.

	Fuzzy App.	Traditional. App.	Number of NE
$S1$	0.81	0.72	815 (288 PER, 348 LOC, 179 ORG)
$S2$	0.87	0.72	783 (280 PER, 331 LOC, 172 ORG)
$S3$	0.85	0.76	627 (217 PER, 267 LOC, 143 ORG)
$S4$	0.77	0.74	587 (218 PER, 235 LOC, 134 ORG)
$S5$	0.92	0.77	259 (94 PER, 115 LOC, 50 ORG)

Table 1. Clustering results.

	Named Entities Cognates
Person	66%
Location	89%
Organization	53%

Table 2. Average percentage of cognate NE identified.

The best results of the F-measure are obtained with the fuzzy approach, so represent the documents by means of different vectors per NE category and combine them with a fuzzy system seems to be a suitable approach for bilingual news clustering. Moreover, as we can see in the Table 2 the average percentage of cognate NE identified is not very high, therefore by improving the cognate NE identification maybe could be possible to obtain better results.

4 Conclusions and Future Work

In this paper we have presented an approach for bilingual news clustering. It is made up of two proposals for two outstanding aspects in MDC: feature selection and document similarity calculation. We represent the documents, the news, only by means of cognate Named Entities (NE). Regarding the similarity, we propose a similarity measure based on a fuzzy rule system. These rules try to incorporate the human knowledge about the importance of the category of the named entities of the news.

The experiments were carried out with five different comparable corpus of news written in English and Spanish, by comparing our approach with a traditional one. The clustering algorithm used belongs to the CLUTO library. Our approach obtained better results with all the corpora, so it seems to be appropriate for bilingual news clustering. The main advantage of using only cognate NE is that no translation resources are needed. On the other hand, the cognate identification approach needs the languages involved in the corpora to be of the same alphabet and linguistic family.

Future work will include the compilation of more news corpora in order to confirm these results and conclusions. We will also explore the application of our approach to other type of documents, such as web pages. We think that the improvement of the cognate identification will also increase the accuracy of the MDC results.

Acknowledgements

References

1. Atserias, J., Casas, B., Comelles, E., González, M., Padró, L. and Padró, M.: Free-Ling 1.3. Syntactic and semantic services in an open-source NLP library. Proceedings of the LREC'06. Genoa, Italy. (2006), http://garraf.epsevg.upc.es/freeling/.

2. Braschler, M., Ripplinger, B. and Schuble, P.: Experiments with the Eurospider Retrieval System for CLEF 2001. LNCS 2406, Springer-Verlag (2002) 102–110.
3. Chau, R., Yeh, C. and Smith, K.A.: A Neural Network Model for Hierarchical Multilingual Text Categorization. Advances in Neural Networks. LNCS 3497, (2005).
4. Chen, H. and Lin, C.: A Multilingual News Summarizer. Proceedings of 18th International Conference on Computational Linguistics (2000) 159–165.
5. Friburger, N. and Maurel, D.: Textual Similarity Based on Proper Names. Mathematical Formal Information Retrieval (MFIR'02), (2002) 155–167.
6. Gang, W.: Named Entity Recognition and An Apply on Document Clustering. *MCSc Thesis*. Dalhousie University, Faculty of Computer Science, Canada, (2004).
7. García-Vega, M., Martínez-Santiago, F., Urea-López, L.A. and Martín-Valdivia, M.T.: Generación de un tesauro de similitud multilinge a partir de un corpus comparable aplicado a CLIR. Procesamiento del Lenguaje Natural, vol. 28, (2002).
8. Gliozzo, A. and Strapparava, C.: 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.
9. Hansen, B.K.: Analog forecasting of ceiling and visibility using fuzzy sets. In Proceedings of the AMS2000 (2000).
10. IPTC - NAA Information Interchange Model Version 4. http://www.iptc.org/std/IIM/4.1/specification/IIMV4.1.pdf.
11. Isermann, R.: On Fuzzy Logic Applications for Automatic Control Supervision and Fault Diagnosis. IEEE Trans.Syst.Man and Cybern, vol 28, (1998) 221–235.
12. Karypis, G.: CLUTO: A Clustering Toolkit. Technical Report: 02-017. University of Minnesota, Department of Computer Science, Minneapolis, MN 55455 (2002).
13. Lawrence, J. L. "Newsblaster Russian-English Clustering Performance Analysis". *Columbia computer science Technical Reports*. http://www1.cs.columbia.edu/ library/2003.html.
14. Mathieu, B., Besancon, R., and Fluhr, C.: Multilingual document clusters discovery. RIAO'2004 (2004) 1–10.
15. Montalvo, S., Martínez, R., Casillas, A. and Fresno, V.: Multilingual Document Clustering: an Heuristic Approach Based on Cognate Named Entities. Proceedings of *COLING-ACL 2006* (2006) 1145–1152.
16. NEClassifier: http://l2r.cs.uiuc.edu/ cogcomp/software.php, (2004).
17. Pouliquen, B., Steinberger, R., Ignat, C., Ksper, E. and Temikova, I.: Multilingual and cross-lingual news topic tracking. Proc. of the CoLing'2004 (2004) 23–27.
18. Rauber, A., Dittenbach, M., Merkl, D.: Towards Automatic Content-Based Organization of Multilingual Digital Libraries: An English, French, and German View of the Russian Information Agency Novosti News. Proceedings of RCDL01 (2001).
19. Silva, J., Mexia, J., Coelho, C. and Lopes, G.: A Statistical Approach for Multilingual Document Clustering and Topic Extraction form Clusters. Pliska Studia Mathematica Bulgarica, **16** (2004) 207–228.
20. Steinberger, R., Pouliquen, B. and Ignat, C.: Exploting multilingual nomenclatures and language-independent text features as an interlingua for cross-lingual text analysis applications. SILTC (2004).
21. Steinberger, R., Pouliquen, B. and Scheer, J.: Cross-Lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC. CICling'2002 (2002).
22. van Rijsbergen, C.J.: Foundations of evaluation. Journal of Documentation vol. 30, (1974) 365–373.