Bilingual Document Clustering using Translation-Independent Features

Claudia Denicia-Carral[1], Manuel Montes-y-Gómez[1],
Luis Villaseñor-Pineda[1] and Rita M. Aceves-Pérez[2]

[1]Laboratory of Language Technologies,
National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico.
{cdenicia, mmontesg, villasen}@inaoep.mx
[2] Department of Electronic Engineering,
Polytechnic University of Altamira (UPALT), Mexico.
rita.aceves@upalt.edu.mx

Abstract. This paper focuses on the task of bilingual clustering, which involves dividing a set of documents from two different languages into a set of thematically homogeneous groups. It mainly proposes a translation independent approach specially suited to deal with linguistically related languages. In particular, it proposes representing the documents by pairs of words orthographically or thematically related. The experimental evaluation in three bilingual collections and using two clustering algorithms demonstrated the appropriateness of the proposed representation, which results are comparable to those from other approaches based on complex linguistic resources such as translation machines, part-of-speech taggers, and named entity recognizers.

1 Introduction

In recent years, due to the globalization phenomenon, there is an increasing interest for organizing and classifying documents from different languages. In this scenario, document clustering aims to identify subsets of documents thematically related in spite of their source language.

The traditional approach for document clustering is based on the assumption that it is possible to establish the topic of documents solely from the frequency of their terms. This basic approach is appropriate for monolingual clustering since all documents may be represented using the same set of words; nevertheless, in a multilingual situation, where documents belong to different languages, it is useless. An immediate solution to this problem is the application of a translation process which allows to construct a common representation for all documents, and, therefore, to apply any existing clustering method.

Even though the translation-based approach is the common strategy for multilingual document clustering (MDC), there are certain linguistically related languages in which it would be possible to apply a translation-independent approach. Particularly, we refer to languages that belong to the same linguistic family (like romance languages), or that by historical reasons or geographic closeness have borrowed a number of words (as the case of Spanish and English). For this kind of languages, it is possible to construct a joint representation of their documents based on words such as common named entities, cognates and foreign words[1].

Taking advantage of the above circumstance, in this paper we explore a translation-independent bilingual clustering approach that represents documents by a set of pairs of related words. We mainly consider two kinds of pairs of related words: on the one hand, orthographically related words such as "presidente-president" or "presidente-presidential", and, on the other hand, thematically related words such as "candidato-voters" or "presidente-elections", which may be extracted from the contexts of the firsts. Therefore, the main contribution of this paper is a method for the extraction of these kinds of pairs of words (herein referred as translation-independent features) and the evaluation of their usefulness as document features in bilingual clustering tasks.

The rest of the paper is organized as follows. Section 2 presents some works on multilingual document clustering. Section 3 details the method for the extraction of translation-independent features. Sections 4 and 5 describe the experimental configuration and results respectively. Finally, Section 6 presents our conclusions and some ideas for future work.

2 Related Work

As we previously mentioned, the translation-based approach is the traditional strategy for MDC. Methods from this approach differentiate one from another by the kind of resources they use for translation as well as by the parts of the texts they translate. There are methods that achieve the translation by means of automatic translation machines [3, 6, 7, 13], and methods that use a bilingual thesaurus or dictionary [12, 14]. Similarly, some of these methods translate the whole documents [6], whereas some others only translate some specific keywords or parts of speech [3, 7, 9, 13].

Motivated by the fact that the performance of this kind of methods is affected by the quality of the automatic translation, Montalvo et al. [8, 9] proposed a translation-independent clustering method that takes advantage from the lexical similarities existing in linguistically related languages. In particular, they proposed using cognate named entities as document features. Their results in a bilingual corpus consisting of documents describing a common set of news events indicate that this kind of features leads to good results in bilingual document clustering.

A possible criticism to the above conclusion might be that it was drawn from a restrictive experimental scenario, where named entities hold a very important role. However, it is expected that for other kind of collections about more general topics, the presence of cognate named entities will be lower, causing the generation of sparse document representations and, therefore, a degradation of the clustering quality. In order to tackle this problem, in this paper we propose to represent documents by a broader set of orthographically similar pairs of words, allowing features such as "presidente-presidential", which are not a translation of each other, but show a clear

[1] Common (or cognate) named entities such as "Barack Obama" which are equally written in Spanish and English; cognates such as "presidente" and "president"; and foreign words such as "software" that is an English word normally used Spanish.

semantic relation. In addition, we propose enriching the representation by including some thematically related pairs of words such as "presidente-elections", which do not present any orthographic similarity, but may be extracted from the contexts of orthographically similar pairs of words.

In order to confirm our claims about the robustness of the proposed features, we present an evaluation that considers three bilingual collections of news reports from the same thematic category but that describe very different events. Somehow, by this experiment, our aim is to investigate the limits of translation-independent features in the task of bilingual document clustering.

3 Extraction of Translation-Independent Features

As we previously mentioned, our proposal is mainly supported on the idea that, for two linguistically related languages, a pair of words having a high orthographic similarity tend to maintain a semantic relation, and, in addition, that the contexts of these words tend to be similar and thematically consistent.

Based on the above assumptions we designed a method for extracting a set of translation-independent features from a given bilingual document collection. This method considers two main steps. The first step focuses on the identification of all orthographically similar pairs of words, whereas, the second uses these pairs of words in order to discover others that tend to co-occur in their contexts, and, therefore, that maintain a "possible" thematic relation.

At the end, we represent the documents from the given bilingual collection using all extracted features, being each feature defined as a pair of related words (w_1, w_2), where w_1 is a word from language L_1 and w_2 is a word from language L_2.

The following two sections describe in detail the extraction of both kinds of features, orthographically and thematically related. Then, Section 3.3 formalizes the representation of documents by the proposed set of features.

3.1 Features based on Orthographic Similarity

Given a document collection (D) containing documents from two different languages (L_1 and L_2), the extraction of this kind of features is carried out as follows:

1. Divide the collection in two sets (D_1 and D_2); each one containing the documents from one single language.
2. Determine the vocabulary (i.e., set of different words) from each language, eliminating the stop words. We mention these sets V_1 and V_2 respectively.
3. Evaluate the orthographic similarity for each pair of words from the two languages; $sim_{ort}(w_i \in V_1,\ w_j \in V_2)$. In our experiments we measured this similarity by the quotient of the length of their longest common subsequence (LCS) and the length of the largest word. For instance, the LCS of the words "*australiano*" (in Spanish) and "*australien*" (in English) is "*a·u·s·t·r·a·l·i·n*", and, therefore, their similarity is 9/11.

4. Select as candidate features all pair of words ($w_i \in V_1$, $w_j \in V_2$) having an orthographic similarity greater than a given specified threshold. That is, we consider that the pair of words (w_i, w_j) is a candidate translation-independent feature if $sim_{ort}(w_i, w_j) \geq \alpha$.
5. Eliminate candidate features (w_i, w_j) that satisfy one of the following conditions: $sim_{ort}(w_i, w_j) < sim_{ort}(w_k \in V_1, w_j)$ or $sim_{ort}(w_i, w_j) < sim_{ort}(w_i, w_k \in V_2)$. The purpose of this final step is to select only the strongest relation for each word, avoiding the generation of many irrelevant features.

At this point it is important to comment that this initial step of our method is similar to other existing approaches for automatic extraction of cognates [2, 5, 10]. It also determines the relation of two words by their orthographic similarity, however, it extracts these pairs of words from the own target document collection avoiding the use of a parallel corpus or bilingual dictionary. Because of this characteristic, the proposed method can extract a great number of related words, some of them incorrect but the vast majority useful for the MDC task. In particular, it may extract pairs of words that are not cognates in a strict sense but that maintain some semantic relation such as “presidencia” (presidency in Spanish) and “president” (in English).

In addition to the extraction of a great number of related pairs of words, this method does not require applying processes for POS tagging or named entity recognition, and, therefore, it may be easily adapted to several pair of languages.

3.2 Features based on Thematic Closeness

As stated in the beginning of Section 3, this second step of the extraction method is based on the idea that the semantic relatedness of two words may be calculated according to their lexical neighbors. Therefore, it considers that a pair of words from different languages ($w_i \in L_1$, $w_j \in L_2$) may be thematically related if they tend to co-occur with the same set of orthographically similar words. In order to illustrate the idea behind the method consider the following example.

Given a bilingual collection formed by documents in Spanish and English, and once extracted a set of orthographically similar features {(presidente, president), (Obama, Obama), …, (congreso, congress)}, it may be possible to assume that the word “elecciones” (elections in Spanish) and “voters” (in English) are thematically related given that “elecciones” tend to co-ocurr with words such “presidente, Obama and congreso”, whereas “voters” co-occur with “president, Obama and congress”.

The following lines describe the general process for the extraction of this kind of features.

Given a collection of documents D with documents written in two different languages, called L_1 and L_2, the extraction of thematically related pairs of words is carried out as follows:

1. Divide the collection in two sets (D_1 and D_2); each one containing the documents from one single language.
2. Determine the vocabulary (i.e., set of different words) from each language, eliminating the stop words. We mention these sets V_1 and V_2 respectively.

3. Select the subset of orthographically “equal” features (E) extracted in the previous step; $E = \{(w_i, w_j) | sim_{ort}(w_i, w_j) = 1\}$.
4. Represent each word from D by a vector $w_i = <p_{i1}, p_{i2},\ldots, p_{i|E|}>$, where p_{ij} indicates the number of documents in which word w_i co-occurs with one of the words from feature j.
5. Compute the similarity for each pair of words from the two languages; $sim_{ocr}(w_i \in V_1, w_j \in V_2)$. In our experiments we measured this similarity based on the vector representations defined in (4) and using the cosine formula.
6. Select as features all pair of words ($w_i \in V_1$, $w_j \in V_2$) having a co-occurrence similarity greater than a given specified threshold. That is, we consider that the pair of words (w_i, w_j) is a translation-independent feature if $sim_{ocr}(w_i, w_j) \geq \beta$.

3.3 Representation of Documents using the Proposed Features

We describe the documents from the bilingual collection D using all extracted features. In particular, we represent each document by a vector $d_i = <p_{i1}, p_{i2},\ldots, p_{i|D|}>$, where p_{ik} indicates the relevance of feature f_k in document d_i. We compute this relevance based on the TF-IDF weighting scheme as indicated below.

Considering that feature f_k is represented by the pair of words (w_{1k}, w_{2k}), where w_{1k} belong to language L_1 and w_{2k} belong to language L_2, p_{ik} is calculated as follows:

$$p_{ik} = TF_{ik} \times IDF_k = \frac{\#\left(w_{xk}, d_i \in D_x\right)}{\left|d_i\right|} \times \log\left(\frac{|D|}{\#\left(w_{1k}, D_1\right) + \#\left(w_{2k}, D_2\right)}\right)$$

where $\#(w_{xk}, d_i)$ indicates the number of occurrences of the word w_{xk} in document d_i, $\#(w_{xk}, D_x)$ the number of documents from language Lx in which w_{xk} occurs, $|d_i|$ the length of document d_i and $|D|$ the number of documents in the whole collection.

4 Experimental Setup

4.1 Evaluation Corpora

The document collection used in the experiments is a selection of news reports from the Reuters Multilingual Corpus Vol. 1 and Vol. 2[2]. This selection includes documents from three languages, namely, Spanish, English and French, and from 16 different categories. Table 2 shows some numbers about this collection.

It is important to remember that all experiments were done using a pair of languages; therefore, we carried out three bilingual experiments: one for Spanish-English considering 922 documents, other for Spanish-French considering 955 documents and another for English-French with 895 documents.

[2] http://trec.nist.gov/data/reuters/reuters.html

Table 1. Corpora Statistics

Language	Documents	Vocabulary without stop words	Words per document (*average*)	Phrases per document (*average*)
Spanish	491	13437	49.19	3.87
English	431	11169	41.06	3.03
French	464	13076	47.34	3.67

4.2 Clustering Algorithms

Given that our aim was to evaluate the usefulness of the proposed features as an individual factor in the task of BDC, we considered a common platform for all experiments, which uses the same weighting scheme for all types of features (TF-IDF), the same similar measure for comparing the documents (cosine measure), as well as two different clustering algorithms.

From the vast diversity of clustering algorithms (for a survey refer to [15]), we decided using the Direct algorithm [4] (a prototype-based approach) and the Star algorithm [1] (a graph-based approach) because:

On the one hand, these algorithms impose different input restrictions; while the first requires knowing the number of desire clusters, the second only needs to consider a minimum threshold (σ) for document similarity.

On the other hand, the Direct algorithm has been previously used in BDC works [8, 9], and the Star algorithm has been recently used in monolingual document clustering tasks [11].

4.3 Evaluation Measure

The used evaluation measure was the *F* measure. This measure allows comparing the automatic clustering solution against a manual clustering (reference solution). It is traditionally computed as described below, where a value of $F = 1$ indicates that the automatic clustering is identical to the manual solution, and a value of $F = 0$ indicates that both solutions do not have any coincidence.

$$F = \sum_{\forall i} \frac{n_i}{n} \max\{F(i,i)\}$$

$$F(i,j) = \frac{2 \times recall(i,j) \times precision(i,j)}{recall(i,j) + precision(i,j)}$$

In this formula, $recall(i,j) = n_{ij}/n_i$ and $precision(i,j) = n_{ij}/n_j$; where n_{ij} is the number of elements of the manual cluster i in the automatic cluster j, n_j is the number of elements of the automatic cluster j and n_i is the number of elements of the manual cluster i.

5 Experimental results

In order to evaluate the appropriateness of the proposed representation we performed three bilingual experiments and considered two different clustering algorithms. Tables 2 and 3 shows the results corresponding to the best experimental configuration indicated by a particular combination of values of α (orthographic similarity threshold), β (co-occurrence similarity threshold) and σ (document similarity threshold for the Star algorithm)[3]. In addition, these tables also include two baseline results: on the one hand, the results achieved by a translation-based method, and, on the other hand, the results from a translation-independent approach using cognate named entities as document features [8, 9]. For the first case we used the translation machine available from Google[4] and applied a document frequency (*DF*) threshold for dimensionality reduction[5] [16], whereas, for the second we performed the recognition of named entities using FreeLing for Spanish, Lingpipe for English and Lia_NE for French[6].

The obtained results show that the proposed method clearly outperforms the approach considering the use of cognate named entities as document features; in average, the MAP scores are 11.6% and 8.6% greater when using the Direct and Star algorithms respectively. From these tables, it is also possible to notice that results from the proposed method are very similar to those from the translation-based method, indicating that our proposal is a competitive alternative when dealing with bilingual collections from linguistically related languages, but having the advantage of not requiring any language processing resource or tool.

Table 2. Results obtained with the Direct clustering algorithm

Languages	Experiment	*F* measure	Best combination
English-Spanish	Using translation	0.21	-
	Using translation (with DF)	0.24	*DF*=5
	Using cognate named entities	0.27	(α = 0.7)
	Using the proposed representation	**0.37**	(α = 0.6; β = 0.9)
French-Spanish	Using translation	0.33	-
	Using translation (with DF)	0.34	*DF*=5
	Using cognate named entities	0.21	(α = 0.7)
	Using the proposed representation	**0.36**	(α = 0.8; β = 0.8)
French-English	Using translation	0.39	-
	Using translation (with DF)	**0.40**	*DF*=5
	Using cognate named entities	0.25	(α = 0.6)
	Using the proposed representation	0.35	(α = 0.7; β = 0.9)

[3] We considered the following values for these thresholds: $\alpha = \{1, 0.9,0.8,0.7,0.6\}$, $\beta = \{1,0.9,0.8\}$, and $\sigma = \{0.1,0.2,0.3,0.4,0.5,0.6\}$.

[4] www.google.com.mx/language_tools

[5] For the experiments we used $DF \geq 1$, $DF \geq 5$ and $DF \geq 10$; the best results were reached using DF $\geq$ *5.*

[6] These tools are available from the following web sites: http://garraf.epsevg.upc.es/freeling/, http://alias-i.com/lingpipe/, http://lia.univ-avignon.fr/.

Table 3. Results obtained with the Star algorithm

Languages	Experiment	*F* measure	Best combination
Spanish-English	Using translation	0.29	($\sigma = 0.1$)
	Using translation (with DF)	**0.30**	($DF = 5$, $\sigma = 0.1$)
	Using cognate named entities	0.25	($\alpha = 0.7$; $\sigma = 0.1$)
	Using the proposed representation	**0.30**	($\alpha = 0.7$; $\beta = 0.9$; $\sigma = 0.1$)
French-Spanish	Using translation	0.25	($\sigma = 0.1$)
	Using translation (with DF)	0.29	($DF = 5$, $\sigma = 0.1$)
	Using cognate named entities	0.21	($\alpha = 0.8$; $\sigma = 0.2$)
	Using the proposed representation	**0.30**	($\alpha = 0.9$; $\beta = 0.9$; $\sigma = 0.1$)
French-English	Using translation	0.27	($\sigma = 0.1$)
	Using translation (with DF)	**0.31**	($DF = 5$, $\sigma = 0.1$)
	Using cognate named entities	0.17	($\alpha = 0.7$; $\sigma = 0.5$)
	Using the proposed representation	0.29	($\alpha = 0.8$; $\beta = 0.9$; $\sigma = 0.2$)

From Tables 2 and 3 it may be argued that the proposed method is sensitive to the selection of the two/three threshold values. In order to clarify the extent of the influence of this selection in the achieved results, Table 5 shows the average and the standard deviation of the *F* measure for all the experiments using the proposed representation and the translation-based approach. These results indicate that the proposed method obtained better average values as well as less standard deviation, allowing to conclude that our method is slightly more robust than the translation-based approach, or, in other words, that all approaches tend to be similarly sensitive to the selection of their parameters.

Table 4. Variability of the results using the Star algorithm (considering all values of α, β and σ for our proposal and $DF = 5$ all values of σ for the translation-based approach)

Language	Experiment	*F* measure	
		Average	*Standard Deviation*
Spanish-English	Translating all to Spanish	0.16	0.08
	Translating all to English	0.17	0.07
	Using the proposed representation	0.19	0.05
French-Spanish	Translating all to Spanish	0.12	0.07
	Translating all to English	0.12	0.07
	Using the proposed representation	0.16	0.06
French-English	Translating all to Spanish	0.15	0.07
	Translating all to English	0.15	0.07
	Using the proposed representation	0.17	0.05

6 Conclusions and Future Work

In this paper we presented a translation-independent bilingual clustering approach that represents documents by a set of pairs of related words. Particularly, we considered two kinds of pairs of related words: orthographically related and thematically related words.

In spite of the complexity of the task –as demonstrated by the achieved results– the representation based on translation independent features shown to be an alternative to the translation-based approach. The results demonstrated that proposed representation is suitable for the clustering task, having the advantage of not depending on any linguistic resource. However, it is important to remember that the application of our proposal is limited to linguistically related languages that belong to the same linguistic family or that by historical reasons or geographic closeness have borrowed a number of words.

Even though the proposed method may be applied to general domain collections, we consider it is more adequate for specific domain document sets, where specialized terms are abundant and tend to be orthographically similar. Regarding this hypothesis, as future work we plan to apply our method to this kind of collections. In addition, we plan to extend the proposed representation to deal with multilingual collections that include documents in more than two languages.

Acknowledgments. This work was done under partial support of CONACYT (Project Grant CB-2007-1-83459 and scholarship 165323).

References

1. Aslam J., Pelekhov, K., and Rus, D. A practical clustering algorithm for static and dynamic information organization. In: *Proceedings of the Symposium on Discrete Algorithms*, 208-217. Washington, D.C., US. 2001.
2. Bergsman, S., Kondrak, G. Multilingual Cognate Identification using Integer Linear Programming. In: *Proceedings of the International Workshop on Acquisition and Management of Multilingual Lexicons*, 11-18. Borovets, Bulgaria. 2007.
3. Chen, H.-H., & Lin, C.-J. A Multilingual News Summarizer. In: *Proceedings of 18th International Conference on Computational Lingüstics*, 159-165. 2000.
4. Karypis, G. CLUTO: A Clustering Toolkit. *Technical Report: 02-017.* University of Minnesota, Department of Computer Science. 2002.
5. Kondrak G. Combining Evidence in Cognate Identification. In: *Proceedings of 17th Conference of the Canadian Society for Computational Studies of Intelligence*, 44-59, London, UK. 2004.
6. Leftin, L. J. Newblaster Russian-English Clustering Performance Analysis. *Technical Report,* Computer Science, Columbia University. 2003.
7. Mathieu, B., Besancon, R., and Fluhr, C. Multilingual Document Clustering Discovery. In *Proceedings of RIAO-04,* Avignon, France, 1-10. 2004.
8. Montalvo, S., Martínez, R., Arantza, C., & Fresno, V. Multilingual News Document Clustering: Two Algorithms Based on Cognate Named Entities. Text, Speech and Dialogue. *Lecture Notes in Artificial Intelligence*, Vol. 4188, 165-172. 2006.
9. Montalvo, S., Martínez, R., Casillas, A., & Fresno, V. Multilingual news clustering: Feature translation vs. identification of cognate named entities. *Pattern Recognition Letters* 28 , 2305-2311. 2007.
10. Mulloni A. and Pekan V. Automatic Detection of Orthografics Cues for Cognate Recognition. In: *Proceedings of LREC*, Genoa, Italy. 2387-2390. 2006.
11. Pérez-Suarez, A., Martinez-Trinidad, F., Carrasco-Ochoa, A., & Medina-Pagola. A New Graph-based Algorithm for Clustering Documents. In: *Proceedings of the*

Foundations of Data Mining workshop (FDM'08), at ICDM'08 Workshops. Italia. 2008.

12. Pouliquen, B., Steinberger, R., Ignat, C., Käsper, E., and Temnikova, I. Multilingual and Cross-lingual News Topic Tracking. In *Proceedings of the 20th International Conference on Computational Linguistics,* 959-965, Geneva, Switzerland. 2004.
13. Rauber, A., Dittenbanch, M., Merkl, D. Towars Automatic Content-based Organization of Multilingual Digital Libraries: an English, French and German View of the Russian Information Agency Novosti News. In: *Third All-Russian Conference Digital Libraries: Advanced Methods and Technologies*, Digital Collections Petrozavodsk. 2001.
14. Steinberger, R., Pouliquen, B., & Hagman, J. Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC. *Lecture Notes in Computer Science*, Vol. 2276, 101-121. 2002.
15. Tan, P. N., Steinbach, M., and Kumar, V. Cluster Analysis: Basic Concepts and Algorithms (chapter 8). In: *Introduction to Data Mining*. Addison-Wesley. 2006.
16. Yang, Y. and Pedersen J. A comparative study on feature selection in text categorization. In: *International Conference on Machine Learning*, 412–420. 1997.