PROJECT MUSE®

Bilingual Document Clustering: Evaluating Cognates as
Features / Le groupage de documents bilingues :
l'évaluation des cognats comme caractéristiques

Claudia Denicia-Carral, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda,
David Pinto-Avendaño

Canadian Journal of Information and Library Science, Volume 35,
Number 3, September/septembre 2011, pp. 265-286 (Article)

Published by University of Toronto Press

DOI: <https://doi.org/10.1353/ils.2011.0022>

➔ *For additional information about this article*

<https://muse.jhu.edu/article/450028>

Bilingual Document Clustering: Evaluating Cognates as Features¹

Le groupage de documents bilingues : l'évaluation des cognats comme caractéristiques

Claudia Denicia-Carral,* Manuel Montes-y-Gómez,* Luis Villaseñor-Pineda,* and David Pinto-Avendaño**

* Laboratory of Language Technologies, Department of Computational Sciences, National Institute of Astrophysics, Optics, and Electronics, Mexico, cdenicia@ccc.inaoep.mx

** Faculty of Computer Science, University of Puebla, Mexico

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 groups, so that documents with similar topics belong to the same group, regardless of their source language. It mainly considers a clustering approach that relies on the use of cognates as document features. Particularly, it proposes two straightforward methods that extract cognates from their own target document collection and do not require using any external bilingual resource, like parallel corpora or a bilingual dictionary. Experimental results in two bilingual collections that include news reports in English and Spanish are encouraging. They indicate that cognates are relevant features for the task of bilingual clustering, outperforming by more than 10% the results achieved by other known approaches.

Keywords: document clustering, multilingualism, bilingual clustering, cognate extraction

Résumé : Cet article se consacre à la tâche du groupage bilingue, qui comprend la répartition d'une série de documents appartenant à deux langues différentes en une série de groupes, de telle façon que les sujets similaires apparaissent dans le même groupe, quelle que soit la langue d'origine. Il s'intéresse surtout à une approche de groupage qui fait usage des cognats considérés comme des traits distinctifs des documents. En particulier, il propose deux méthodes directes permettant l'extraction des cognats à partir de leur propre collection de documents cibles, sans recourir à l'utilisation de ressources bilingues externes, telles que des corpus parallèles ou un dictionnaire bilingue. Nous avons obtenu des résultats expérimentaux encourageants avec deux collections bilingues incluant des bulletins de nouvelles en anglais et en espagnol. Ces résultats indiquent que les cognats sont des traits distinctifs valables pour le groupage de documents

bilingues, et qu'ils permettent d'obtenir des résultats dépassant de 10 % ceux que l'on obtient avec les autres approches connues.

Mots-clés : groupage de documents, multilinguisme, groupage bilingue, extraction de cognats

Introduction

Great advances in communication and storage media provide more information than ever before. This information can satisfy almost every information need; nevertheless, without the appropriate management facilities, all of it is practically useless. This fact has motivated the emergence of several information retrieval (IR) methods that help in searching and analysing large document collections (Baeza-Yates and Ribeiro-Neto 1999).

At present, due to the Internet explosion and the existence of several multicultural communities, one of the major challenges that face this kind of methods is multilingualism. In particular, in a multilingual scenario, it is expected that an IR method could retrieve information written in languages different from that of the user's query (Grefenstette 1998).

Research in multilingual IR has shown great advances in the last decade. Researchers have proposed different translation procedures, indexing schemes, weighting approaches, and information-fusion methods (refer to the TREC² and CLEF³ conferences). Nevertheless, in the application of text-mining techniques within IR, it is interesting to notice that while there are many successful examples of monolingual IR, such as the use of association rules for query expansion (Song et al. 2007; Wei, Bresson, and Ooi 2000) and the use of clustering techniques for the visualization of search results (Cigarran et al. 2005; Leuski and Croft 1996; Zeng et al. 2004), there have been only a few attempts to apply these techniques to multilingual IR. We understand the main reason is the lack of a vast set of text-mining techniques to deal with information expressed in a mixture of languages.

To tackle this problem, it is necessary to design more techniques specially suited to multilingual text mining. In line with this purpose, in this paper we focus on the problem of bilingual document clustering (BDC),

which involves dividing a set of documents, written in two different languages, into a set of groups or clusters, so that documents with similar topics belong to the same group, regardless of their source language (Montalvo et al. 2006).

Evidently, traditional clustering strategies cannot be directly applied in BDC, since they require that all documents be represented using the same set of features (i.e., the same set of words). As a result of this constraint, the most popular approach for BDC considers the use of translation technologies. The idea is to construct, by means of a translation procedure, a common representation for all documents, and then apply a clustering algorithm. This approach has achieved satisfactory results but has also revealed its dependence on the availability and quality of translation machines and other bilingual resources.

As an alternative approach, to reduce dependence on multilingual resources, we consider a clustering method that relies on translation-independent features. In particular, we use cognates as document features. Cognates, as defined in Kondrak, Marcu, and Knight (2003), are words in different languages that are similar in their orthographic or phonetic form and that are possible translations of each other, such as the words *presidente* (in Spanish) and *president* (in English). This feature limits the application of this approach to languages with the same alphabet and belonging to the same linguistic family (like romance languages), or that have borrowed a number of words through history or geographic closeness (as the case of Spanish and English).

As expected, a key task of this approach is the extraction of cognates. We propose to use two different methods, which differ from traditional approaches (Bergsman and Kondrak 2007; Kondrak 2004; Mulloni and Pekan 2006; Ribeiro et al. 2001) in that they extract the cognates from their own target document collection and do not require external bilingual resources such as parallel corpora or a bilingual dictionary. In addition, they differ from similar approaches in that they are not limited to extracting certain types of cognates, such as is the case for Montalvo et al. (2007), which centred on the extraction of noun cognates and cognate-named entities.

To evaluate the proposed methods we carried out experiments in two different bilingual collections that include news reports in English and Spanish. The experimental results are encouraging; they indicate that

cognates are relevant features for BDC, outperforming the results from a translation-based approach as well as those achieved by using other kinds of translation-independent features such as cognate-named entities.

The rest of the paper is organized as follows. Section 2 presents previous work in multilingual document clustering. Section 3 describes two methods for extracting cognates from a given bilingual document collection. Section 4 shows the experimental results and compares them against those from other BDC approaches. Section 5 gives a deep analysis of our results, and finally, section 6 presents our conclusions and describes future work directions.

Related work

As we previously mentioned, the most popular approach for BDC is based on the use of translation procedures. Methods from this approach can be categorized from two different perspectives: according to the kinds of resources used for the translation, and according to the parts of texts that are translated.

In particular, some methods achieve the translation by using automatic translation machines (Hsin-Hsi and Lin 2000; Leftin 2003; Mathieu, Bescanson, and Fluhr 2004; Rauber, Dittenbanch, and Merkl 2001), whereas some others use a bilingual thesaurus or dictionary (Pouliquen et al. 2004; Steinberger, Pouliquen, and Hagman 2002). Similarly, some of these methods translate the whole document (Leftin 2003), while others only translate specific keywords or parts of speech (Hsin-Hsi and Li 2000; Mathieu, Bescanson, and Fluhr 2004; Montalvo et al. 2007; Rauber, Dittenbanch, and Merkl 2001).

Methods using the translation-based approach tend to obtain satisfactory results, especially for general-domain collections. However, they present drawbacks: some depend on the availability and quality of translation machines, while others rely on the existence and extension of bilingual resources such as dictionaries and thesaurus, and most of them are greatly affected by the semantic ambiguity of words.

More recently, Montalvo et al. (2006) proposed an alternative approach based on the use of translation-independent features. In particular, they used cognate-named entities as document features. They evaluated these

features in a comparable corpus⁴ consisting of clusters with small granularity (formed by news reports of the same event) and demonstrated their appropriateness for BDC. Subsequently, Montalvo et al. (2007) compared the results achieved by cognate-named entities against those from noun cognates and translation-based features and concluded that “the use of named entities as the only type of features to represent news leads to good BDC results.”

A possible criticism of this conclusion is that it was drawn from a restrictive experimental scenario, where named entities hold a very important role. Our hypothesis is that in collections with clusters of greater granularity, the presence of cognate-named entities will be lower, causing the generation of sparse document representations and, therefore, a degradation of the clustering quality. To tackle this problem, in this paper we propose constructing richer document representations by considering a greater number of features. In particular, we propose using (all kinds of) cognates as features instead of using only cognate-named entities or noun cognates.

To confirm our claims about the robustness of the proposed features, we present an extended evaluation that considers not only the same corpus used by Montalvo et al., but also an additional comparable corpus organized in clusters that include news reports corresponding to the same thematic category but that describe very different events. By this experiment, our aim is to investigate the limits of translation-independent features in the task of bilingual document clustering.

Automatic extraction of cognates

As we stated in the previous section, cognates are words in different languages that are similar in their orthographic or phonetic form and that are possible translations of each other. As a result of their usefulness in machine translation, their extraction has received great attention in recent years. In particular, most methods extract cognates from parallel corpora or bilingual dictionaries by measuring the orthographic similarity of words having a common meaning (Bergsman and Kondrak 2007; Kondrak et al. 2003; Mulloni and Pekan 2006; Ribeiro et al. 2001).

In this section, we introduce two new methods for the extraction of cognates. These methods, different from previous approaches that aim

to detect all cognates from a pair of languages, focus exclusively on the extraction of cognates from a given bilingual document collection. As a result of this restriction, and motivated by their application in diverse collections (from different languages and concerning different topics), we designed these methods to be very simple and general, avoiding the use of any external bilingual resource and complex procedures for natural language processing.

Global extraction of cognates

This first method is very general since it does not assume any information about the nature of the bilingual corpus that will be clustered. It applies mainly an orthographic similarity measure over the two vocabularies of the corpus, and determines a set of features from a global perspective, without considering their contexts.

As expected, this method allows extracting many features (cognates), but, because of its simplicity, it also generates a lot of false friends, that is, pairs of words that are orthographically similar but have different meanings; for instance, the Spanish word *subir* (“to raise”) and the French word *subir* (“to suffer”). Nonetheless, under the assumption that the occurrence of false friends is less than that of true cognates, we decided to maintain the generality and simplicity of this method, and exclude the implementation of any action for their elimination.

The following paragraphs describe the main steps of this method.

Given a document collection (D) containing documents from two different languages (L_1 and L_2), the extraction of cognates from D 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 or set of different words from each language (V_1 and V_2 respectively). For this step we eliminate the stopwords.
3. Evaluate the orthographic similarity for each pair of words from the two languages; $\text{sim}(w_i \in V_1, w_j \in V_2)$. This similarity is calculated as the quotient of the length of the longest common subsequence

(LCS) of the two words and the length of the largest word. For example, the LCS of the words *australiano* (in Spanish) and *Australian* (in English) is “a u s t r a l i a n”; therefore, their orthographic similarity is 10/11.

4. Select as candidate cognates 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, $w_i \in V_1$ and $w_j \in V_2$ will be defined as cognates if $\text{sim}(w_i \in V_1, w_j \in V_2) \geq \beta$.
5. Eliminate the candidate cognates ($w_i \in V_1, w_j \in V_2$) that satisfy one of the following conditions (where a better candidate cognate w_k is found for a given word):
 - a. $\text{sim}(w_i, w_j) < \text{sim}(w_k \in V_1, w_j)$
 - b. $\text{sim}(w_i, w_j) < \text{sim}(w_i, w_k \in V_2)$

Local extraction of cognates

As mentioned above, as a result of its simplicity, the first method might extract several features describing pairs of words that, even though they are orthographically similar, they do not maintain a semantic relation. On the basis of this observation, this second method aims to enhance the precision of the cognate extraction by emulating the use of a parallel corpus. Mainly, it is supported on the assumption that the contexts of a given named entity are very similar in both languages, and, therefore, that the words occurring in these contexts tend to have similar meanings. See, for instance, the contexts of the named entity “*Lino Oviedo*” in table 1.

This second method considers, in a first step, the identification of common named entities (cognate-named entities), and, in a second step, the extraction of cognates exclusively from these contexts. In the following we describe this method in detail.

Given a document collection (D) containing documents from two different languages (L_1 and L_2), the extraction of cognates from D 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.

Table 1: Example contexts of the named entity "Lino Oviedo" in both languages

Entity	Language	Context
Lino Oviedo	Spanish	La Policía paraguaya desconocía esta noche el supuesto arresto del ex general golpista Lino Oviedo en la zona de la Triple Frontera, donde convergen los límites de Argentina, Brasil y Paraguay, informaron a EFE fuentes oficiales.
	English	Former Paraguayan coup leader General Lino Oviedo, who was arrested over the weekend by Brazilian authorities, was moved from Foz do Iguazu to Brasilia on Monday, federal police told EFE.

2. Locate the set of named entities from each sub-collection (E_1 and E_2 respectively). In particular, we consider the identification of named entities denoting persons, organizations, and locations (details of the named entity extractor are given below).
3. Align similar named entities. We consider that two named entities $e_i \in E_1$ and $e_j \in E_2$ are the same if their orthographic similarity is greater than a given specified threshold: that is, if $\text{sim}(e_i \in E_1, e_j \in E_2) \geq \beta$.
4. For each aligned pair of named entities ($e_i \in E_1, e_j \in E_2$), perform the following steps:
 - a. Extract all sentences from D_1 containing e_i and all sentences from D_2 including e_j . We call these subsets of sentences S_1 and S_2 respectively.
 - b. Determine the "local" vocabularies V_1 and V_2 from S_1 and S_2 .
 - c. Evaluate the orthographic similarity of each pair of words from the local vocabularies: $\text{sim}(w_i \in V_1, w_j \in V_2)$. For this evaluation we used the LCS measure.
 - d. Select as candidate cognates all pairs of words ($w_i \in V_1, w_j \in V_2$) having an orthographic similarity greater than a given specified threshold. That is, $w_i \in V_1$ and $w_j \in V_2$ will be defined as cognates if $\text{sim}(w_i \in V_1, w_j \in V_2) \geq \beta$.
 - e. Include as candidate cognates all pairs of words from the aligned entities. That is, if $e_i = "w_{i1} w_{i2} \dots w_{in}"$ and $e_j = "w_{j1} w_{j2} \dots w_{jn}"$, then the pairs of words (w_{i1}, w_{j1}), (w_{i2}, w_{j2}), and so on) will be also considered cognates.

5. From the complete set of candidate cognates, obtained from all aligned pairs of entities, eliminate the candidate cognates ($w_i \in L_1$, $w_j \in L_2$) that satisfy one of the following two conditions, as above:
 - f. $\text{sim}(w_i, w_j) < \text{sim}(w_k \in L_1, w_j)$
 - g. $\text{sim}(w_i, w_j) < \text{sim}(w_i, w_k \in L_2)$

Experimental evaluation

This section describes the evaluation of the use of cognates as document features. First, it presents the experimental setup; it mainly describes the evaluation corpora as well as the evaluation measure. Then it presents some baseline results. Specifically, it shows the results achieved by a translation-based approach as well as by the usage of cognate-named entities as document features. Finally, it presents the clustering results corresponding to the use of the globally and locally extracted cognates.

Experimental setup

Evaluation corpora

For the experiments, we used two different bilingual collections that include news reports in English and Spanish. The first collection, called *UNED corpus* (Montalvo et al. 2006; 2007), is a set of news reports provided by the EFE agency and compiled through the HERMES project.⁵ This corpus consists of 192 news reports, 100 in Spanish and 92 in English, organized in 35 groups, 33 multilingual and 2 monolingual. The second collection, called *RCV corpus*, is a selection of documents from the Reuters Multilingual Corpus, volumes 1 and 2.⁶ It consists of 922 news reports, 491 in Spanish and 431 in English, distributed in 16 multilingual groups. Table 2 shows some numbers from these collections.

Language resources

As we previously mentioned, the proposed methods for cognate extraction do not make use of any bilingual resource (parallel corpus or bilingual dictionary). However, the local extraction of cognates relies on the identification of named entities. For that purpose we employed two different named entity recognizers: FreeLing⁷ for Spanish, and Lingpipe⁸ for English.

Table 2: News reports in English and Spanish

Corpus	Language	Vocabulary Size	Average Words per Document	Average Sentences per Document
UNED	Spanish	5,645	90.39	4.40
	English	3,595	53.37	3.66
RCV	Spanish	13,437	49.19	3.87
	English	11,169	41.06	3.03

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 Tan, Steinbach, and Kumar 2006), we decided to use the Direct algorithm (a prototype-based approach) (Karypis 2002) and the Star algorithm (a graph-based approach) (Aslam, Pelekov, and Rus 2004) for two reasons:

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

Second, the Direct algorithm has been previously used in BDC works (Montalvo et al. 2006; 2007), and the Star algorithm has been recently used in monolingual document clustering tasks (Suárez et al. 2008).

Evaluation measure

The evaluation measure used was the F measure, which allows comparison of the automatic clustering solution against a manual clustering (reference solution). It is traditionally computed as described below,⁹ 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 are completely different.

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

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

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 .

Results

Baseline results

We selected as a baseline the results from two different approaches. First, the results achieved by a translation-based approach,¹⁰ translating all documents to the second language, and second, the results corresponding to the use of cognate-named entities as document features. Tables 3 and 4 show the results from both approaches.

It is important to comment that these tables (as well as subsequent ones) indicate only the best result from each experiment. These results were achieved by a specific configuration, indicated by a particular selection of

Table 3: Baseline results for the UNED corpus

Experiment Description	Clustering Algorithm	Best Configuration	F Measure
Translating all news to English	Star	($\sigma = 0.1$)	0.781
Translating all news to Spanish	Star	($\sigma = 0.1$)	0.787
Using cognate-named entities	Star	($\beta = 0.8$; $\sigma = 0.5$)	0.835
Translating all news to English	Direct	–	0.853
Translating all news to Spanish	Direct	–	0.869
Using cognate-named entities	Direct	($\beta = 0.7$)	0.842

Table 4: Baseline results for the RCV corpus

Experiment Description	Clustering Algorithm	Best Configuration	F Measure
Translating all news to English	Star	($\sigma = 0.1$)	0.292
Translating all news to Spanish	Star	($\sigma = 0.1$)	0.279
Using cognate-named entities	Star	($\beta = 0.7$; $\sigma = 0.3$)	0.248
Translating all news to English	Direct	–	0.211
Translating all news to Spanish	Direct	–	0.207
Using cognate-named entities	Direct	($\beta = 0.7$)	0.265

the values of β and σ , which denote the orthographic similarity threshold and the document similarity threshold respectively.¹¹

From these tables it is interesting to notice that results from the translation-based approach are very stable; they are almost the same from both configurations, translating all documents to English or to Spanish. In addition, it is also interesting to notice that in the first case (using the UNED corpus) all approaches were helped by knowledge about the number of groups, whereas using the RCV corpus, only the one based on translation-independent features was favoured. We presume this behaviour was a consequence of the nature of collections. On the one hand, the UNED corpus consists of groups related to very specific subjects (i.e., clusters of small granularity) such as “the visit of Clinton to China” or “presidential elections in Mexico,” whereas on the other hand, the RCV corpus includes groups of very general topics such as economy, politics, and sports (i.e., clusters of greater granularity).

Results of the proposed representation

As we previously mentioned, one motivation for our proposal was the sparse representations generated by cognate-named entity features. It is well known that having poor document representations generates several problems in the clustering procedure, since there are insufficient elements to correctly evaluate the similarity between documents. Table 5 shows that the two proposed methods allowed generating more features than using the cognate-named entities exclusively. This table also shows that when using these new sets of features (globally and locally extracted cognates) it was possible to represent all documents from the given collection.¹²

Table 5: Some numbers from different kinds features

		UNED		RCV	
Type of Features	β	Number of Features	Unrepresented Documents	Number of Features	Non-represented Documents
Cognate-named entities	1.0	164	6	327	164
	0.9	159	6	339	155
	0.8	176	4	445	74
	0.7	201	2	665	27
	0.6	272	0	1154	6
Globally extracted cognates	1.0	333	0	790	18
	0.9	412	0	948	10
	0.8	866	0	2279	5
	0.7	1415	0	3863	0
	0.6	1950	0	5166	0
Locally extracted cognates	1.0	291	0	652	27
	0.9	320	0	693	19
	0.8	475	0	1058	8
	0.7	662	0	1509	5
	0.6	1084	0	2326	1

Tables 6 and 7 show the clustering results achieved by the two proposed sets of features. These results demonstrate the pertinence of these features in the task of bilingual document clustering. In particular, they show that:

- The performance of both kinds of features (globally and locally extracted cognates) is almost independent from the selected clustering algorithm. In other words, the difference between the F measures achieved by both kinds of features was practically the same when using the Star or Direct clustering algorithms. In particular, in the UNED corpus, locally extracted cognates outperformed globally extracted features by approximately 3% using any clustering algorithm, whereas in the RCV corpus, globally extracted cognates outperformed locally extracted features by approximately 1%.

Table 6: Results of the proposed features in the UNED corpus

Experiment Description	Clustering Algorithm	Best Configuration	F measure
Using cognate-named entities	Star	$(\beta = 0.8; \sigma = 0.5)$	0.835
Using globally extracted cognates	Star	$(\beta = 0.8; \sigma = 0.3)$	0.838
Using locally extracted cognates	Star	$(\beta = 0.8; \sigma = 0.3)$	0.863
Using cognate-named entities	Direct	$(\beta = 0.7)$	0.842
Using globally extracted cognates	Direct	$(\beta = 0.8)$	0.908
Using locally extracted cognates	Direct	$(\beta = 1.0)$	0.932

Table 7: Results of the proposed features in the RCV corpus

Experiment Description	Clustering Algorithm	Best Configuration	F Measure
Using cognate-named entities	Star	$(\beta = 0.7; \sigma = 0.3)$	0.248
Using globally extracted cognates	Star	$(\beta = 0.7; \sigma = 0.1)$	0.276
Using locally extracted cognates	Star	$(\beta = 0.7; \sigma = 0.2)$	0.265
Using cognate-named entities	Direct	$(\beta = 0.7)$	0.265
Using globally extracted cognates	Direct	$(\beta = 0.8)$	0.317
Using locally extracted cognates	Direct	$(\beta = 0.6)$	0.293

- The use of more features improves the representation of documents. This is evidenced by the fact that the proposed features achieve better results than those from cognate-named entities when applying the same clustering algorithm. As additional evidence, these results also outperform those reported by Montalvo et al. (2007), where cognate-named entities and noun cognates were extracted using a different approach.
- The results achieved by globally and locally extracted cognates are similar. However, we presume that locally extracted cognates would be useful only for collections formed by clusters with small granularity, where a significant number of cognate-named entities may exist, such as the case of the UNED corpus.
- The obtained results are comparable to (and in most cases slightly better than) those from the translation-based approach (refer to tables 3 and 4). Nevertheless, the major advantage of our proposal is that it does not rely on the use of any bilingual resource.

Analysis of results

In order to have a better understanding of the achieved results, we performed an evaluation of the “relative hardness” of both bilingual corpora. For this evaluation, we adopted some ideas for measuring the cluster validity via correlation (Tan et al. 2006). In particular, given the similarity matrix for a corpus as well as the cluster labels from its documents (according to a manual clustering), we evaluated the relative hardness of the corpus by looking at the correlation between the similarity matrix and an ideal version of the similarity matrix based on the cluster labels. In an ideal similarity matrix, documents belonging to the same cluster have a similarity of 1, and documents from different cluster have a similarity of 0. Therefore, a high correlation between the ideal and actual similarity matrices indicates that documents belonging to the same cluster are close to each other (low hardness), while low correlation indicates the opposite (high hardness).

Table 8 shows the correlation values from both corpora obtained by using different kinds of features.¹³ These results show that it is more difficult to discover the structure of RCV corpus than that from the UNED corpus using any kind of features.

On the other hand, given that the similarity of documents depends directly on their representation, these results also provide evidence that, for collections formed by low-granularity clusters, cognate-named entities may lead to good clustering results, as concluded by Montalvo et al. (2007). However, for collections with clusters of greater granularity, cognate-named entities are not enough, and globally extracted cognates represent a better alternative.

Among the techniques used for measuring the cluster validity via correlation, there is also a (qualitative) visual approach. In this approach, rows

Table 8: Quantitative results about the relative hardness of the used corpora

Representation	UNED Corpus	RCV Corpus
Using cognate-named entities	0.868	0.178
Using globally extracted cognates	0.833	0.237
Using locally extracted cognates	0.879	0.179

and columns from the similarity matrix are sorted so that all documents belonging to the same group are placed together. In theory, using this distribution, if the corpus has well-separated clusters, then the similarity matrix should be roughly block-diagonal. If not, the patterns displayed in the similarity matrix can reveal the relationships between clusters.

Figures 1 and 2 show the plotting of the similarity matrices from the UNED and RCV corpora using the three kinds of translation-independent features.¹⁴

The figures confirm that the relative hardness of the RCV corpus is greater than that of the UNED corpus, independent of the kind of used features. Visually speaking, we can observe that all the representations of the UNED corpus generate clusters around the diagonal, while the plotting of the RCV corpus is completely different. In this case, we can observe that most of its documents share similarities not only with documents from the same group but also with documents belonging to other clusters.

Additionally, the proposed features produce less noisy graphs than those generated by cognate-named entities, which, to some extent, explains the best *F* measures of our proposals (refer to tables 6 and 7). This fact was especially evident for the RCV corpus, where globally extracted cognates (see figure 2b) allowed eliminating most of the inter-cluster similarities.

Conclusions and future work

In this paper we describe a bilingual clustering approach that considers the use of translation-independent features. In particular, we propose the use of cognates as document features.

The experimental results demonstrate that using cognates extracted by straightforward methods as document features is an adequate strategy for bilingual document clustering. In particular, using the Star algorithm, our clustering approach could outperform the traditional translation-based approach by 10.2% and the results from cognate-named entity features by 11.3%. On the other hand, when using the Direct algorithm (having knowledge about the number of groups in the manual solution), our method outperformed the other approaches by 46.9% and 19.6% respectively.

(a) Cognate-named entities

(b) Globally extracted cognates

(c) Locally extracted cognates

Figure 1: Relative hardness of the UNED corpus using different features

(a) Cognate-named entities

(b) Globally extracted cognates

(c) Locally extracted cognates

Figure 2: Relative hardness of the RCV corpus using different features

In this paper we also introduced two different methods for the automatic extraction of cognates. These methods differ from previous approaches in that they extract the cognates from their own target document collection and do not require using external bilingual resource like parallel corpora or a bilingual dictionary. As a result of this characteristic, we consider that these methods are general enough to be applied in different domains, especially in technical ones, where specialized terminology is shared across languages. However, future work must confirm this claim. On the other hand, the main restriction of these methods is that they are applicable only to languages having the same alphabet and belonging to the same linguistic family (e.g., the romance languages).

It is also important to comment that our results indicate that the appropriate method for cognate extraction is determined by inherent characteristics of the corpus to be clustered. That is, using locally extracted cognates seems more appropriate for a corpus formed by groups of small granularity (about very particular topics), whereas globally extracted cognates tend to be better for a corpus with clusters of greater granularity (of general thematic categories).

Finally, it is clear that there is a long road ahead for this task, especially concerning the definition of better bilingual document representations. Given that, we plan to focus our future work on defining other kinds of translation-independent features. It is clear that cognates on their own are not sufficient to describe the content of documents from a bilingual collection. Therefore, we propose to enrich them by adding a set of bilingual associated words, such as *politics/ministro*, *Obama/secretario*, etc., extracted from the contexts of known cognates such as *president/presidente*.

Notes

1. This work was done under partial support from CONACYT-Mexico (project grant 83459 and scholarship 165323). We also want to thank Soto Montalvo and Raquel Martinez from UNED-Spain for the resources provided.
2. Text REtrieval Conferences. See <http://www.trec.nist.gov/>.
3. Cross-Language Evaluation Forum. See <http://www.clef-campaign.org/>.
4. While a parallel corpus contains texts and their translations, a comparable corpus consists of sets of similar texts in different languages that are not translations of each other (Bowker and Pearson 2002).
5. Hermes, <http://nlp.uned.es/hermes/index.html>.
6. Reuters Corpora, <http://trec.nist.gov/data/reuters/reuters.html>.

7. <http://nlp.lsi.upc.edu/freeling/>.
8. LingPipe, <http://alias-i.com/lingpipe/>.
9. This formula is commonly applied to evaluate clustering results, and it is an adaptation of the classical F measure used for IR evaluation. Details on this measure can be found in Tan, Steinbach, and Kumar (2006).
10. For this experiment we used the translation machine available from Google, at http://www.google.com.mx/language_tools.
11. Remember that the orthographic similarity was computed by means of the longest common subsequence (LCS) measure and the document similarity by the cosine measure. In all experiments we used $\beta = \{1, 0.9, 0.8, 0.7, 0.6\}$ and $\sigma = \{0.1, 0.2, 0.3, 0.4, 0.5\}$.
12. In table 5, the column "Unrepresented documents" indicates the number of documents that had none of the used features, that is, the number of documents having all their vector values equal to zero.
13. In this case, the values of β were kept as those shown in tables 6 and 7.
14. The figures show only the points corresponding to similarities greater than the values of σ shown at tables 6 and 7. These figures were computed using the WaCOS web-based system. See <http://nlp.cs.buap.mx/watermarker/>.

References

- Aslam, J., K. Pelehov, and D. Rus. 2004. The star clustering algorithm for static and dynamic information organization. *Journal of Graph Algorithms and Applications* 8, no. 1: 95–129.
- Baeza-Yates, R., and B. Ribeiro-Nieto. 1999. *Modern information retrieval*. Wokingham, UK: Addison-Wesley.
- Bergsman, S., and G. Kondrak. 2007. Multilingual cognate identification using integer linear programming. In *Proceedings of the International Workshop on Acquisition and Management of Multilingual Lexicons*, 11–18. Borovets, Bulgaria.
- Bowker, L., and J. Pearson. 2002. *Working with specialized language: A practical guide to using corpora*. London: Routledge.
- Cigarran, J., A. Peñas, J. Gonzalo, and F. Verdejo. 2005. Evaluating hierarchical clustering of search results. *Lecture Notes in Computer Science* 3772: 49–54.
- Grefenstette, G. 1998. *Cross-language information retrieval*. Norwell, MA: Kluwer Academic Publishers.
- Hsin-Hsi, C., and C-J Lin. 2000. Multilingual news summarizer. In *Proceedings of 18th International Conference on Computational Linguistics*, 159–65. Stroudsburg, PA: Association for Computational Linguistics.
- Karypis, G. 2002. *CLUTO: A clustering toolkit*. Technical Report 02-017. Department of Computer Science, University of Minnesota, Minneapolis, MN.
- Kondrak, G. 2004. Combining evidence in cognate identification. In *Proceedings of Canadian AI 2004: Lecture Notes in Computer Science* 3060: 44–59. London: Springer.

- Kondrak, G., D. Marcu, and K. Knight. 2003. Cognates can improve statistical translation models. In *Proceedings of NAACL*, 46–8. Stroudsburg, PA: Association for Computational Linguistics: Edmonton.
- Leftin, L.J. 2003. *Newblaster Russian-English clustering performance analysis*. Technical report. Computer Science, Columbia University.
- Leuski, A., and B. Croft. 1996. *An evaluation of techniques for clustering search results*. Technical Report IR-76. Department of Computer Science, University of Massachusetts, Amherst.
- Mathieu, B., R. Besancon, and C. Fluhr. 2004. Multilingual document clustering discovery. In *Proceedings of RIAO-04, Avignon, France*, ed. Christian Fluhr, Gregory Grefenstette, and W. Bruce Croft, 1–10. Avignon: University of Avignon.
- Montalvo, S., R. Martínez, C. Arantza, and V. Fresno. 2006. Multilingual news document clustering: Two algorithms based on cognate named entities. *Lecture Notes in Artificial Intelligence* 4188: 165–72.
- Montalvo, S., R. Martínez, A. Casillas, and V. Fresno. 2007. Multilingual news clustering: Feature translation vs. identification of cognate named entities. *Pattern Recognition Letters* 28: 2305–11.
- Mulloni, A., and V. Pekan. 2006. Automatic detection of orthographics cues for cognate recognition. In *Proceedings of LREC, Genoa*. 2387–90. Paris: European Language Resources Association.
- Pouliquen, B., R. Steinberger, C. Ignat, E. Käsper, and I. Temnikova. 2004. Multilingual and cross-lingual news topic tracking. In *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, 959–65. Stroudsburg, PA: Association for Computational Linguistics.
- Rauber, A., M. Dittenbanch, and D. Merkl. 2001. Towards 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*, 88–95.
- Ribeiro, A., G. Dias, G. Lopes, and J. Mexia. 2001. Cognates alignment. In *Proceedings of the Machine Translation Summit VIII (MT Summit VIII)—Machine Translation in the Information Age*. Santiago de Compostela, Spain. 287–92.
- Song, M., I.-Y. Song, X. Hu, and R. Allen. 2007. Integration of association rules and ontologies for semantic query expansion. *Data & Knowledge Engineering* 63(1): 63–75.
- Steinberger, R., B. Pouliquen, and J. Hagman. 2002. *Cross-lingual document similarity calculation using the multilingual thesaurus EUROVOC*. In *Proceedings of CICLing 2002, Mexico City, Mexico: Lecture Notes in Computer Science* 2276: 415–24. London: Springer.
- Suárez, A.P., J.F.M. Trinidad, J.A.C. Ochoa, and J.E.M. Pagola. 2008. A new graph-based algorithm for clustering documents. In *Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW '08), Pisa, Italy*, 710–19. Piscataway, NJ: IEEE.

- Tan, P.N., M. Steinbach, and V. Kumar. 2006. Cluster analysis: Basic concepts and algorithms. In *Introduction to data mining*, chap. 8. Boston: Addison-Wesley.
- Wei, J., S. Bressan, and B. Ooi. 2000. Mining term association rules for automatic global query expansion: Methodology and preliminary results. In *Proceedings of the First International Conference on Web Information Systems Engineering (WISE'00)*, 1: 366–73. Hong Kong: IEEE.
- Zeng, H., Q. He, Z. Chen, W. Ma, and J. Ma. 2004. Learning to cluster web search results. In *Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval*, 210–14. Sheffield: ACM.