Multilingual Document Clustering: an Heuristic Approach Based on Cognate Named Entities

Soto Montalvo
GAVAB Group
URJC
soto.montalvo@urjc.es

Raquel Martínez
NLP&IR Group
UNED
raquel@lsi.uned.es

Arantza Casillas
Dpt. EE
UPV-EHU
arantza.casillas@ehu.es

Víctor Fresno
GAVAB Group
URJC
victor.fresno@urjc.es

Abstract

This paper presents an approach for Multilingual Document Clustering in comparable corpora. The algorithm is of heuristic nature and it uses 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 bilingual or multilingual resources. However, it depends on the possibility of identifying cognate named entities between the languages used in the corpus. An additional advantage of the approach is that it does not need any information about the right number of clusters; the algorithm calculates it. We have tested this approach with a comparable corpus of news written in English and Spanish. In addition, we have compared the results with a system which translates selected document features. The obtained results are encouraging.

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 the documents that are similar to other documents are 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. 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 of documents or clustering it could not be so relevant and even it could be a source of noise.

In this work we dealt with MDC and we proposed an approach based only on cognate Named Entities (NE) identification. We have tested this approach with a comparable corpus of news written in English and Spanish, obtaining encouraging results. 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

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 1145–1152,
Sydney, July 2006.

to be provided to the algorithm. It 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.

In order to compare the results of our approach with other based on features translation, we also dealt with this one, as baseline approach. The system uses EuroWordNet (Vossen, 1998) to translate the features. We tried different features categories and combinations of them in order to determine which ones lead to improve MDC results in this approach.

In the following section we relate previous work in the field. In Section 3 we present our approach for MDC. Section 4 describes the system we compare our approach with, as well as the experiments and the results. Finally, Section 5 summarizes the conclusions and the future work.

2 Related Work

MDC is normally applied with parallel (Silva et. al., 2004) or comparable corpus (Chen and Lin, 2000), (Rauber et. al., 2001), (Lawrence, 2003), (Steinberger et. al., 2002), (Mathieu et. al, 2004), (Pouliquen et. al., 2004). 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: (1) translate the whole document to an anchor language, and (2) translate some features of the document to an anchor language.

With regard to the first approach, some authors use machine translation systems, whereas others translate the document word by word consulting a bilingual dictionary. In (Lawrence, 2003), the author presents several experiments for clustering a Russian-English multilingual corpus; several of these experiments are based on using a machine translation system. Columbia's Newsblaster system (Kirk et al., 2004) clusters news into events, it categorizes events into broad topic and summarizes multiple articles on each event. In the clustering process non-English documents are translated using simple dictionary lookup techniques for translating Japanese and Russian documents, and the Systran translation system for the other languages used in the system.

When the solution involves translating only some features, first it is necessary to select these features (usually entities, verbs, nouns) and then translate them with a bilingual dictionary or/and consulting a parallel corpus.

In (Mathieu et. al, 2004) before the clustering process, the authors perform a linguistic analysis which extracts lemmas and recognizes named entities (location, organization, person, time expression, numeric expression, product or event); then, the documents are represented by a set of terms (*keywords* or named entity types). In addition, they use document frequency to select relevant features among the extracted terms. Finally, the solution uses bilingual dictionaries to translate the selected features. In (Rauber et. al., 2001) the authors present a methodology in which documents are parsed to extract features: all the words which appear in n documents except the stopwords. Then, standard machine translation techniques are used to create a monolingual corpus. After the translation process the documents are automatically organized into separate clusters using an un-supervised neural network.

Some approaches first carry out an independent clustering in each language, that is a monolingual clustering, and then they find relations among the obtained clusters generating the multilingual clusters. Others solutions start with a multilingual clustering to look for relations between the documents of all the involved languages. This is the case of (Chen and Lin, 2000), where the authors propose an architecture of multilingual news summarizer which includes monolingual and multilingual clustering; the multilingual clustering takes input from the monolingual clusters. The authors select different type of features depending on the clustering: for the monolingual clustering they use only named entities, for the multilingual clustering they extract verbs besides named entities.

The strategies that use language-independent representation try to normalize or standardize the document contents in a language-neutral way; for example: (1) by mapping text contents to an independent knowledge representation, or (2) 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. Normally, in a multilingual thesaurus, elements in different languages are related via language-independent items. So, two documents written in different languages can be considered similar if they have similar representation according to the thesaurus. In some cases, it is necessary to use the thesaurus in combination with a machine learning method for mapping correctly documents onto thesaurus. In (Steinberger et. al., 2002) the authors present an approach to calculate the semantic similarity by representing the document contents in a language independent way, using the descriptor terms of the multilingual thesaurus *Eurovoc*.

The second approach, recognition of language independent text features, involves the recognition of elements such as: dates, numbers, and named entities. In others works, for instance (Silva et. al., 2004), the authors present a method based on Relevant Expressions (RE). The RE are multilingual lexical units of any length automatically extracted from the documents using the LiPXtractor extractor, a language independent statistics-based tool. The RE are used as base features to obtain a reduced set of new features for the multilingual clustering, but the clusters obtained are monolingual.

Others works combine recognition of independent text features (numbers, dates, names, cognates) with mapping text contents to a thesaurus. In (Pouliquen et. al., 2004) the cross-lingual news cluster similarity is based on a linear combination of three types of input: (a) cognates, (b) automatically detected references of geographical place names, and (c) the results of a mapping process onto a multilingual classification system which maps documents onto the multilingual thesaurus *Eurovoc*. In (Steinberger et. al., 2004) it is proposed to extract language-independent text features using gazetteers and regular expressions besides thesaurus and classification systems.

None of the revised works use as unique evidence for multilingual clustering the identification of cognate named entities between both sides of the comparable corpora.

3 MDC by Cognate NE Identification

We propose an approach for MDC based only on cognate NE identification. The NEs categories that we take into account are: PERSON, ORGANIZATION, LOCATION, and MISCELLANY. Other numerical categories such as DATE, TIME or NUMBER are not considered because 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 process has two main phases: (1) cognate NE identification and (2) clustering. Both phases are described in detail in the following sections.

3.1 Cognate NE identification

This phase consists of three steps:

1. Detection and classification of the NEs in each side of the corpus.
2. Identification of cognates between the NEs of both sides of the comparable corpus.
3. To work out a statistic of the number of documents that share cognates of the different NE categories.

Regarding the first step, it is carried out in each side of the corpus separately. In our case we used a corpus with morphosyntactical annotations and the NEs identified and classified with the FreeLing tool (Carreras et al., 2004).

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

- Obtaining two list of NEs, one for each language.
- Identification of entity mentions in each language. 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 Levenshtein 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.

Finally, a statistic of the number of documents that share cognates of the different NE categories is worked out. This information can be used by the algorithm (or the user) to select the NE category used as constraint in the clustering steps 1(a) and 2(b).

3.2 Clustering

The algorithm for clustering multilingual documents based on cognate NEs is of heuristic nature. It consists of 3 main phases: (1) first clusters creation, (2) addition of remaining documents to existing clusters, and (3) final cluster adjustment.

1. First clusters creation. This phase consists of 2 steps.

   (a) First, documents in different languages that have more cognates in common than a threshold are grouped into the same cluster. In addition, at least one of the cognates has to be of a specific category (PERSON, LOCATION or ORGANIZATION), and the number of mentions has to be similar; a threshold determines the similarity degree. After this step some documents are assigned to clusters while the others are free (with no cluster assigned).

   (b) Next, it is tried to assign each free document to an existing cluster. This is possible if there is a document in the cluster that has more cognates in common with the free document than a threshold, with no constraints regarding the NE category. If it is not possible, a new cluster is created. This step can also have as result free documents.

   At this point the number of clusters created is fixed for the next phase.

2. Addition of the rest of the documents to existing clusters. This phase is carried out in 2 steps.

   (a) A document is added to a cluster that contains a document which has more cognates in common than a threshold.

   (b) Until now, the cognate NEs have been compared between both sides of the corpus, that is a bilingual comparison. In this step, the NEs of a language are compared with those of the same language. This can be described like a monolingual comparison step. The aim is to group similar documents of the same language if the bilingual comparison steps have not been successful. As in the other cases, a document is added to a cluster with at least a document of the same language which has more cognates in common than a threshold. In addition, at least one of the cognates have to be of a specific category (PERSON, LOCATION or ORGANIZATION).

3. Final cluster adjustment. Finally, if there are still free documents, each one is assigned to the cluster with more cognates in common, without constraints or threshold. Nonetheless, if free documents are left because they do not have any cognates in common with those assigned to the existing clusters, new clusters can be created.

Most of the thresholds 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. For example, in steps 1(a) and 2(b) we enforce at least on match in a specific NE category. This parameter can be customized in order to guide the grouping towards some type of NE. In Section 4.5 the exact values we used are described.

Our approach is an heuristic method that following an agglomerative approach and in an iterative way, decides the number of clusters and locates each document in a cluster; everything is based in cognate NEs identification. The final number of clusters depends on the threshold values.

4 Evaluation

We wanted not only determine whether our approach was successful for MDC or not, but we also wanted to compare its results with other approach based on feature translation. That is why we try MDC by selecting and translating the features of the documents.

In this Section, first the MCD by feature translation is described; next, the corpus, the experiments and the results are presented.

4.1 MDC by Feature Translation

In this approach we emphasize the feature selection based on NEs identification and the grammatical category of the words. The selection of features we applied is based on previous work (Casillas et. al, 2004), in which several document representations are tested in order to study which of them lead to better monolingual clustering results. We used this MDC approach as baseline method.

The approach we implemented consists of the following steps:

1. Selection of features (NE, noun, verb, adjective, ...) and its context (the whole document or the first paragraph). Normally, the journalist style includes the heart of the news in the first paragraph; taking this into account we have experimented with the whole document and only with the first paragraph.

2. Translation of the features by using EuroWordNet 1.0. We translate English into Spanish. When more than one sense for a single word is provided, we disambiguate by selecting one sense if it appears in the Spanish corpus. Since we work with a comparable corpus, we expect that the correct translation of a word appears in it.

3. In order to generate the document representation we use the TF-IDF function to weight the features.

4. Use of an clustering algorithm. Particularly, we used a partitioning algorithm of the CLUTO (Karypis, 2002) library for clustering.

4.2 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 have used three subset of that collection. The first subset, call $S1$, consists on 65 news, 32 in Spanish and 33 in English; we used it in order to train the threshold values. The second one, $S2$, is composed of 79 Spanish news and 70 English news, that is 149 news. The third subset, $S3$, contains 179 news: 93 in Spanish and 86 in English.

In order to test the MDC results we carried out a manual clustering with each subset. Three persons read every document and grouped them considering the content of each one. They judged independently and only the identical resultant clusters were selected. The human clustering solution is composed of 12 clusters for subset $S1$, 26 clusters for subset $S2$, and 33 clusters for $S3$. All the clusters are multilingual in the three subsets.

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

4.3 Evaluation metric

The quality of the experimentation results are determined by means of an external evaluation measure, the F-measure (van Rijsbergen, 1974). 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.

4.4 Experiments and Results with MDC by Feature Translation

After trying with features of different grammatical categories and combinations of them, Table 1 and Table 2 only show the best results of the experiments.

The first column of both tables indicates the features used in clustering: NOM (nouns), VER (verbs), ADJ (adjectives), ALL (all the lemmas), NE (named entities), and 1^{rst} PAR (those of the first paragraph of the previous categories). The second column is the F-measure, and the third one indicates the number of multilingual clusters obtained. Note that the number of total clusters of each subset is provided to the clustering algorithm. As can be seen in the tables, the results depend on the features selected.

4.5 Experiments and Results with MDC by Cognate NE

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.

Regarding the thresholds of the clustering phase (Section 3.2), after training the thresholds with the collection $S1$ of 65 news articles we have concluded:

- The first step in the clustering phase, 1(a), performs a good first grouping with threshold relatively high; in this case 6 or 7. That is, documents in different languages that have more cognates in common than 6 or 7 are grouped into the same cluster. In addition, at least one of the cognates have to be of an specific category, and the difference between the number of mentions have to be equal or less than 2. Of course, these threshold are applied after checking that there are documents that meet the requirements. If they do not, thresholds are reduced. This first step creates multilingual clusters with high cohesiveness.

- Steps 1(b) and 2(a) lead to good results with small threshold values: 1 or 2. They are designed to give priority to the addition of documents to existing clusters. In fact, only step 1(b) can create new clusters.

- Step 2(b) tries to group similar documents of the same language when the bilingual comparison steps could not be able to deal with them. This step leads to good results with a threshold value similar to 1(a) step, and with the same NE category.

On the other hand, regarding the NE category enforce on match in steps 1(a) and 2(b), we tried with the two NE categories of cognates shared by the most number of documents. Particularly, with $S2$ and $S3$ corpus the NE categories of the cognates shared by the most number of documents was LOCATION followed by PERSON. We experimented with both categories.

Table 3 and Table 4 show the results of the application of the cognate NE approach to subsets $S2$ and $S3$ respectively. The first column of both tables indicates the thresholds for each step of the algorithm. Second and third columns show the results by selecting PERSON category as NE category to be shared by at least a cognate in steps 1(a) and 2(b); whereas fourth and fifth columns are calculated with LOCATION NE category. The results are quite similar but slightly better with LOCATION category, that is the cognate NE category shared by the most number of documents. Although none of the results got the exact number of clusters, it is remarkable that the resulting values are close to the right ones. In fact, no information about the right number of cluster is provided to the algorithm.

If we compare the performance of the two approaches (Table 3 with Table 1 and Table 4 with Table 2) our approach obtains better results. With the subset $S3$ the results of the F-measure of both approaches are more similar than with the subset $S2$, but the F-measure values of our approach are still slightly better.

To sum up, our approach obtains slightly better results that the one based on feature translation with the same corpora. In addition, the number of multilingual clusters is closer to the reference solution. We think that it is remarkable that our approach reaches results that can be comparable with those obtained by means of features translation. We will have to test the algorithm with different corpora (with some monolingual clusters, different languages) in order to confirm its performance.

5 Conclusions and Future Work

We have presented a novel approach for Multilingual Document Clustering based only on cognate

Selected Features	**F-measure**	**Multilin. Clus./Total**
NOM, VER	0.8533	21/26
NOM, ADJ	0.8405	21/26
ALL	0.8209	21/26
NE	0.8117	19/26
NOM, VER, ADJ	0.7984	20/26
NOM, VER, ADJ, 1^{rst} PAR	0.7570	21/26
NOM, ADJ, 1^{rst} PAR	0.7515	22/26
ALL, 1^{rst} PAR	0.7473	19/26
NOM, VER, 1^{rst} PAR	0.7371	20/26

Table 1: MDC results with the feature translation approach and subset $S2$

Selected Features	**F-measure**	**Multilin. Clus. /Total**
NOM, ADJ	0.8291	26/33
ALL	0.8126	27/33
NOM, VER	0.8028	26/33
NE	0.8015	23/33
NOM, VER, ADJ	0.7917	25/33
NOM, ADJ, 1^{rst} PAR	0.7520	28/33
NOM, VER, ADJ, 1^{rst} PAR	0.7484	26/33
ALL, 1^{rst} PAR	0.7288	26/33
NOM, VER, 1^{rst} PAR	0.7200	24/33

Table 2: MDC results with the feature translation approach and subset $S3$

Thresholds				**1(a), 2(b) match on PERSON**		**1(a), 2(b) match on LOCATION**	
Steps				**Results**	**Clusters**	**Results**	**Clusters**
1(a)	**1(b)**	**2(a)**	**2(b)**	**F-measure**	**Multil./Calc./Total**	**F-measure**	**Multil./Calc./Total**
6	2	1	5	**0.9097**	24/24/26	**0.9097**	24/24/26
6	2	1	6	0.8961	24/24/26	0.8961	24/24/26
6	2	1	7	0.8955	24/24/26	0.8955	24/24/26
6	2	2	5	0.8861	24/24/26	0.8913	24/24/26
7	2	1	5	0.8859	24/24/26	0.8913	24/24/26
6	2	2	4	0.8785	24/24/26	0.8899	24/24/26
6	2	2	6	0.8773	24/24/26	0.8833	24/24/26
6	2	2	7	0.8773	24/24/26	0.8708	24/24/26

Table 3: MDC results with the cognate NE approach and $S2$ subset

Thresholds				**1(a), 2(b) match on PERSON**		**1(a), 2(b) match on LOCATION**	
Steps				**Results**	**Clusters**	**Results**	**Clusters**
1(a)	**1(b)**	**2(a)**	**2(b)**	**F-measure**	**Multil./Calc./Total**	**F-measure**	**Multil./Calc./Total**
7	2	1	5	**0.8587**	30/30/33	**0.8621**	30/30/33
6	2	1	5	0.8552	30/30/33	0.8552	30/30/33
6	2	1	6	0.8482	30/30/33	0.8483	30/30/33
6	2	1	7	0.8471	30/30/33	0.8470	30/30/33
6	2	2	5	0.8354	30/30/33	0.8393	30/30/33
6	2	2	6	0.8353	30/30/33	0.8474	30/30/33
6	2	2	4	0.8323	30/30/33	0.8474	30/30/33
6	2	2	7	0.8213	30/30/33	0.8134	30/30/33

Table 4: MDC results with the cognate NE approach and $S3$ subset

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 permit 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 by using the threshold values of the algorithm.

We have tested this approach with a comparable corpus of news written in English and Spanish, obtaining encouraging results. We think that 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. In addition, we have compared our approach with other based on feature translation, resulting that our approach presents a slightly better performance.

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. In addition, we will study if changing the order of the bilingual and monolingual comparison steps the performance varies significantly 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

Benoit Mathieu, Romanic Besancon and Christian Fluhr. 2004. "Multilingual document clusters discovery". RIAO'2004, p. 1-10.

Arantza Casillas, M. Teresa González de Lena and Raquel Martínez. 2004. "Sampling and Feature Selection in a Genetic Algorithm for Document Clustering". *Computational Linguistics and Intelligent Text Processing*, CICLing'04. Lecture Notes in Computer Science, Springer-Verlag, p. 601-612.

Hsin-Hsi Chen and Chuan-Jie Lin. 2000. "A Multilingual News Summarizer". *Proceedings of 18th International Conference on Computational Linguistics*, p. 159-165.

Xavier Carreras, I. Chao, Lluis Padró and M. Padró 2004 "An Open-Source Suite of Language Analyzers". *Proceedings of the 4th International Conference on Language Resources and Evaluation* (LREC'04). Lisbon, Portugal. http://garraf.epsevg.upc.es/freeling/.

Karypis G. 2002. " CLUTO: A Clustering Toolkit". *Technical Report: 02-017*. University of Minnesota, Department of Computer Science, Minneapolis, MN 55455.

David Kirk Evans, Judith L. Klavans and Kathleen McKeown. 2004. "Columbian Newsblaster: Multilingual News Summarization on the Web". *Proceedings of the Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting*, HLT-NAACL'2004.

Lawrence J. Leftin. 2003. "Newsblaster Russian-English Clustering Performance Analysis". *Columbia computer science Technical Reports*.

Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, Emilia Ksper and Irina Temikova. 2004. "Multilingual and cross-lingual news topic tracking". *Proceedings of the* 20^{th} *International Conference on computational Linguistics*, p. 23-27.

Andreas Rauber, Michael Dittenbach and Dieter Merkl. 2001. "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.

van Rijsbergen, C.J. 1974. "Foundations of evaluation". *Journal of Documentation*, 30 (1974), p. 365-373.

Joaquin Silva, J. Mexia, Carlos Coelho and Gabriel Lopes. 2004. "A Statistical Approach for Multilingual Document Clustering and Topic Extraction form Clusters". *Pliska Studia Mathematica Bulgarica*, v.16,p. 207-228.

Ralf Steinberger, Bruno Pouliquen, and Johan Scheer. 2002. "Cross-Lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC". *Computational Linguistics and Intelligent Text Processing*, CICling'02. Lecture Notes in Computer Science, Springer-Verlag, p. 415-424.

Ralf Steinberger, Bruno Pouliquen, and Camelia Ignat. 2004. "Exploiting multilingual nomenclatures and language-independent text features as an interlingua for cross-lingual text analysis applications". *Slovenian Language Technology Conference. Information Society*, SLTC 2004.

Vossen, P. 1998. "Introduction to EuroWordNet". *Computers and the Humanities Special Issue on EuroWordNet*.