Procesamiento de Lenguaje Natural, Revista nº 48 marzo de 2012, pp 81-88 recibido 16-01-12 revisado 06-03-12 aceptado 18-03-12

NESM: a Named Entity based Proximity Measure for Multilingual News Clustering*

NESM: una medida de similitud para el Clustering Multilingüe de Noticias basada en Entidades Nombradas

Soto Montalvo
Dpto. Ciencias de la Computación
Universidad Rey Juan Carlos
soto.montalvo@urjc.es

Víctor Fresno, Raquel Martínez
Dpto. Lenguajes y Sistemas Informáticos
UNED
{vfresno, raquel}@lsi.uned.es

Resumen: Una de las tareas esenciales dentro del proceso del *Clustering* de Documentos es medir la similitud entre éstos. En este trabajo se presenta una nueva medida basada en el número y la categoría de las Entidades Nombradas compartidas entre documentos. Para evaluar la calidad de la medida propuesta en el clustering multilingüe de noticias, se han utilizado tres medidas de pesado diferentes y dos medidas de similitud estándar. Los resultados demuestran, con tres colecciones de noticias comparables escritas en español e inglés, que la medida propuesta es competitiva, superando en algunos casos a medidas como el coseno y el coeficiente de correlación.
Palabras clave: Entidad Nombrada, *Clustering* Multilingüe, Similitud de documentos

Abstract: Measuring the similarity between documents is an essential task in Document Clustering. This paper presents a new metric that is based on the number and the category of the Named Entities shared between news documents. Three different feature-weighting functions and two standard similarity measures were used to evaluate the quality of the proposed measure in multilingual news clustering. The results, with three different collections of comparable news written in English and Spanish, indicate that the new metric performance is in some cases better than standard similarity measures such as cosine similarity and correlation coefficient.
Keywords: Named Entity, Multilingual Clustering, Document Similarity

1. Introduction

Multilingual Document Clustering (MDC) involves dividing a set of n documents, written in different languages, into k number of clusters; such that similar documents belong to the same cluster. A multilingual cluster contains documents written in different languages; and a monolingual cluster is formed by documents written in the same language. The scope of MDC tools is to ease tasks such as multilingual information access and organization: Cross-Lingual Information Retrieval or Question Answering in multilingual collections, among others.

Multilingual Document Clustering systems have developed different solutions to group related documents. Mainly two groups of strategies have been employed: (1) those that use translation technologies and (2) those that transform the document into a language-independent representation. The approaches that use Machine Translation systems, such as (Flaounas et al., 2011), (Lawrence, 2003), (Gael and Zhu, 2007), or those that use translation with dictionaries, such as (Cheung, Huang, and Lam, 2004), (Urizar and Loinaz, 2007), (Mathieu, Besançon, and Fluhr, 2004), (Wu and Lu, 2007), belong to the first strategy. On the other hand, the approaches that map text contents to an independent knowledge representation, such as thesaurus (Steinberger, Pouliquen, and Hagman, 2002), (Steinberger, Pouliquen, and Ignat, 2006), (Pouliquen et al., 2004); or those that recognize language-independent text features inside the documents, such as (Denicia-Carral et al., 2010), (Steinberger,

* This work has been part-funded by the Education Council of the Regional Government of Madrid, MA2VICMR (S-2009/TIC-1542), and the research project Holopedia, funded by the Ministerio de Ciencia e Innovación under grant TIN2010-21128-C02.

ISSN 1135-5948

Pouliquen, and Ignat, 2005), (Silva et al., 2004), (Chau, Yeh, and Smith, 2005), belong to the second one. Both strategies can be either used isolated or combined. We are at work on an approach that employs the latter, representing the document contents by using the cognate Named Entities (NEs) as language-independent features.

Named Entities play a critical role in Natural Language Processing (NLP) and Information Retrieval (IR) tasks, such as Machine Translation, Document Clustering, Summarization, Information Extraction, etc. Particularly, NEs are more informative than other features in news documents, as we can be seen in different works: (Shinyama and Sekine, 2004), (Shah, Croft, and Jensen, 2006), (Pouliquen et al., 2004), (Armour, Japkowicz, and Matwin, 2005), and (Montalvo et al., 2007b). According to the jargon of journalists, the content of a news document must contain answers to six questions (*What*, *Who*, *When*, *Where*, *Why*, *How*), and part of the answers of these questions usually involve NEs (e.g., answers to *who* questions usually involve persons or organizations, answers to *where* questions involve locations, and answers to *when* questions involve temporal expressions). For this reason our proposal to measure the similarity between documents are based on the shared NEs.

There are several factors that influence Document Clustering results such as the clustering algorithm, the features that represent the documents, the feature-weighting function, and the similarity measure. In particular, many problems that involve some type of document organization depend on the estimated similarity, or distance, between them. Finding the similarity between documents is usually based on extracting features from the documents, weighting those features, and using standard functions such as the cosine measure or the correlation coefficient. Thus, a clustering algorithm that exploits special characteristics of the document content may lead to superior results (Kogan, Teboulle, and Nicholas, 2005). A large number of functions that estimate similarity (or distance) between documents have been developed, varying greatly in their expressiveness, mathematical properties, or assumptions (Rodríguez, 2002), (Baeza-Yates and Ribeiro-Neto, 1999). Hence, the calculation of the similarity can differ depending on the particular domain, corpus, features, or task. In this paper, we focus on the news similarity calculation, presenting a new measure to compare documents with the aim of improving Multilingual News Clustering. This new measure determines the similarity between multilingual documents using information from the cognate NEs that they contain.

The rest of the paper is organized as follows. Section 2 presents the proposed measure. Section 3 shows how news documents are represented, as well as the clustering algorithm used in the experimentation. The experimental evaluation is presented in Section 4, and finally, Section 5 presents the conclusions and future work.

2. A new comparison measure: NESM

Since we are exploring news clustering based on the representation of the news by means of the cognate NEs, we wanted to check the impact and behaviour of a measure which explicitly takes into account the number and categories of the different NEs shared between news. With this aim we propose the Named Entities Shared Measure (NESM) that is defined as follows:

$$NESM(\mathbf{d_1}, \mathbf{d_2}) = \sum_{NE\ cat} \frac{NE(\mathbf{d_1}, \mathbf{d_2})_{sha}}{NE(\mathbf{d_1}, \mathbf{d_2})_{max}} \quad (1)$$

Where $NE(\mathbf{d_1}, \mathbf{d_2})_{sha}$ is the number of different NEs shared by two documents d_1 and d_2, and $NE(\mathbf{d_1}, \mathbf{d_2})_{max}$ is the maximum number of different NEs shared by two documents of the corpus written in the same languages as d_1 and d_2. For example, in a corpus where documents are written in Spanish and English, if d_1 and d_2 are written in Spanish, $NE(\mathbf{d_1}, \mathbf{d_2})_{max}$ is the maximum number of NEs shared by two documents in the Spanish corpus side; if d_1 and d_2 are written in English, the maximum value is the maximum number of NEs shared by two documents in the English corpus side; and, finally, if d_1 and d_2 are written one in Spanish and the other in English, the maximum value is the maximum number of NEs shared by two documents written in Spanish and English in the corpus.

NESM is not formally a similarity measure because it does not fulfill all the required metric properties. Even so, this measure is well-behaved mathematically. In detail, NESM is

not enclosed in [0,1], but [0,4] because we take into account 4 NE categories. Furthermore, due to its special normalization it can be that $NESM(\mathbf{d_i}, \mathbf{d_i}) \neq 1$. This is because we are dividing by a maximum value in the corpus. With this normalization we try to find out how close the number of NEs shared by two documents is to the maximum, and therefore this measure does not keep the main required condition for a similarity measure. So, NESM function is a proximity measure between documents. The higher the value of NESM the more similar are the documents.

3. Multilingual News Clustering

Next, we describe the way the documents are represented in our approach, the functions we use to compare news in addition to NESM, and the clustering algorithm we use in these experiments.

3.1. News Representation

Usually news documents contain a large number of NEs. The frequency of named-entity phrases in news texts reflects the significance of the events they are associated with (Kumaran and Allan, 2004). Named Entities tend to be preserved across comparable documents because it is generally difficult to paraphrase names (Shinyama and Sekine, 2004). Taking into account this synchronicity of names in comparable news texts, our approach for Multilingual News Clustering is based on the representation of the news by means of the cognate NEs they contain. In linguistics, cognates are words that have a common etymological origin and share common spelling. In (Montalvo et al., 2007a), the authors showed that the use of cognate NEs, as the only type of features to represent news, leads to good multilingual clustering performance. The results were comparable, and in some cases even better, to those obtained by using more types of features (nouns, verbs,...) with the feature translation approach.

In our approach the cognate NEs identification consists of three steps: (1) detection and classification of the NEs in each side of the corpus (each monolingual corpus); (2) identification of cognate NEs; and (3) working out a statistic of the number of documents that share cognates of the different NE categories.

Regarding the first step, the NE detection and classification, it is carried out in each monolingual corpus separately using available NE detection and classification tools. In section 4.1 we describe the tools used.

In order to identify the cognate NEs, second step, the following phases are carried out. First, we obtain lists of NEs, one for each language. Next, we identify entity mentions in each language. Then, the identification of cognate NEs between the different sides of the comparable corpus is carried out. The identification of the cognates, as well as the identification of the entity mentions, are based on the use of the Levensthein edit-distance function (Levenshtein, 1966). As a result we obtain a list of cognate NEs. With all this information, the statistic of the third step is worked out.

In order to calculate the list of cognate NEs we take into account their specific category as well. We only consider the following NE categories: PERSON, ORGANIZATION, LOCATION and MISCELLANY, since they can be suitable to find common content in documents in a multilingual news corpus. Other categories, such as DATE, TIME or NUMBER are not taken into account.

To evaluate the NESM function we use a representation of the news based on the presence or absence of shared NEs between the documents, considering their categories. The representation for the other measures we use in the experiments is based on the vector space model (Salton, 1983). In this model a document is represented through a vector, where each component represents the weight of a feature in the document. In this case each component represents a NE of the collection vocabulary, and the component value reflects the importance of the NE in the news text. We have compared NESM with the following term-weighting functions:

- **Binary (Bin)**. The weight of a feature t in a document vector $\mathbf{d}$ is given by: $B(t, \mathbf{d}) = \{0, 1\}$, which is a binary function that represents if the document d contains the feature t. If d contains t, the value is 1, otherwise is 0.
- **Term Frequency (TF)**. Each term or feature is assumed to have importance proportional to the number of times it occurs in the document. The weight of a feature t in a document vector $\mathbf{d}$ is given by: $W(t, \mathbf{d}) = TF(t, \mathbf{d})$, where $TF(t, \mathbf{d})$ is the frequency of the feature t in the

document d.

- **TF-IDF**. It is the combination of TF and IDF to weight terms. The combination weight of a feature t in a document vector $\mathbf{d}$ is given by: $W(t, \mathbf{d}) = TF(t, \mathbf{d}) \times IDF(t)$. The IDF factor of a feature t is given by: $IDF(t) = log\frac{N}{df(t)}$, where N is the number of documents in the collection and $df(t)$ is the number of documents that contain the feature t.

3.2. Baseline Similarity Functions

We use as baseline similarity functions two standard functions such as the cosine similarity (COS) and the correlation coefficient (CORR). Then we calculate the distance between the vectors that represent the news documents using these two functions and NESM, the measure we present.

The COS and CORR measures are well known in literature. A popular measure of similarity for text clustering (which normalizes the features by the covariance matrix) is the cosine of the angle between two vectors. The cosine measure is given by

$$COS(\mathbf{d}_1, \mathbf{d}_2) = \frac{\mathbf{d}_1 \mathbf{d}_2}{\sqrt{\sum_i d_{1i}^2} \cdot \sqrt{\sum_i d_{2i}^2}} \quad (2)$$

A strong property is that the cosine similarity does not depend on the length: $COS(\alpha \mathbf{d}_1, \mathbf{d}_2) = COS(\mathbf{d}_1, \mathbf{d}_2)$ for $\alpha > 0$. This makes it the most popular measure for text documents. Also, due to this property, samples can be normalized to the unit sphere for more efficient processing (Dhillon and Modha, 2001).

The correlation coefficient is often used to predict a feature from a highly similar group of objects whose features are known. The Pearson correlation is defined as

$$CORR(\mathbf{d}_1, \mathbf{d}_2) = \frac{(\mathbf{d}_1 - \bar{d_1}) \cdot (\mathbf{d}_2 - \bar{d_2})}{\sqrt{\sum_i (d_{1i} - \bar{d_1})^2 \cdot \sum_i (d_{2i} - \bar{d_2})^2}} \quad (3)$$

where $\bar{d}$ denotes the average feature value of $\mathbf{d}$ over all dimensions.

3.3. Clustering Algorithm

In these experiments we use the "Agglomerative" clustering algorithm from the well known CLUTO library (Karypis, 2003). A similarity matrix and a specific number of clusters are needed by the algorithm. When NESM measure is considered, the proximity matrix is used as a similarity matrix. In connection with the number of clusters, we use the number of the reference solution.

4. Experimental Evaluation

This Section describes the test environment: the news collections, the evaluation measure used, and the results.

4.1. Document Collections

We carried out the experiments with several comparable corpora of news, written in Spanish and English, that come from three different sources: S1, S2 and S3.

S1 is a compilation of news from the news agency EFE and the same period of time, compiled by the HERMES project[1]. Three persons, independently, read every document and manually grouped them considering the content of each one. S2 is a subset of CLEF-2003 (Savoy, 2003) collection of news. In this case we use the category label of the news to build the reference solution. The data sets from this collection have more documents per cluster than those from the other ones, and also they have more monolingual clusters. S3 is a compilation of news downloaded from the webs of different newspapers: *El Mundo* and *El País* (in Spanish); *The Guardian*, *BBC News*, *The Daily Telegraph*, *Washington Post*, and *New York Times* (in English). We have used a crawler system that selects from the international news of *El Mundo*[2] the "related links" in order to create clusters according to a topic. The grouping proposed by the system was revised and corrected by three persons independently.

We performed a linguistic analysis of each document of the three collections, by means of the Freeling tool (Carreras et al., 2004). Specifically we carried out: morphosyntactic analysis, lemmatization, and recognition and classification of NEs. The Named Entity Tagger Software (Ratinov and Roth, 2009) is used to detect and classify the NEs of the English documents.

We randomly generated 13 data sets for each collection. The data sets have different sizes, and most of them are non-uniformly distributed per category. S2 and S3 have several monolingual clusters, whereas S1 collection has mainly multilingual clusters. Ta-

[1] http://nlp.uned.es/hermes/index.html

[2] www.elmundo.es

ble 1 provides a description of the data sets, where: the first column identifies the data set; the second one shows the number of documents in Spanish and English, respectively; the third column provides the number of multilingual and monolingual clusters of the reference solution; the fourth one shows the average number of NEs per document; and finally, the five column contains the average number of documents per cluster in the reference solution.

	Docs ES-EN	Clusters Mul-Mon	Avg. NEs/D	Avg. D/Clust
S1DS1	12-12	3-2	7.75	4.8
S1DS2	21-19	6-2	8.57	5
S1DS3	25-22	9-1	8.61	4.7
S1DS4	33-32	13-2	8.29	4.3
S1DS5	37-34	13-4	8.05	4.1
S1DS6	43-41	16-1	8.15	4.9
S1DS7	48-47	18-2	8.18	4.7
S1DS8	58-56	18-5	8.68	4.9
S1DS9	60-60	20-3	9.11	5.2
S1DS10	64-64	23-2	9.08	5.1
S1DS11	78-78	28-2	9.79	5.2
S1DS12	81-81	29-2	9.77	5.2
S1DS13	100-92	33-2	10.39	5.4
S2DS1	10-9	1-2	4.21	6.3
S2DS2	15-14	1-3	6.93	7.2
S2DS3	19-19	1-3	8.89	9.5
S2DS4	31-31	2-2	10.87	15.5
S2DS5	35-35	2-2	11.81	17.5
S2DS6	40-40	2-2	12.43	20
S2DS7	46-46	2-2	13.15	23
S2DS8	51-50	2-3	13.20	20.2
S2DS9	52-51	2-3	13.13	20.6
S2DS10	59-58	2-4	13.17	19.5
S2DS11	72-70	2-6	13.33	17.7
S2DS12	80-78	3-5	13.06	19.7
S2DS13	110-109	4-7	13.05	20
S3DS1	6-7	1-1	22.53	6.5
S3DS2	8-8	1-2	22.12	5.3
S3DS3	17-14	1-6	19.48	4.4
S3DS4	25-24	4-6	20.97	4.9
S3DS5	28-27	4-7	20.89	5
S3DS6	38-36	5-8	22.64	5.6
S3DS7	42-39	5-9	23.74	5.7
S3DS8	60-56	5-10	25.18	7.7
S3DS9	64-60	5-10	25.08	8.2
S3DS10	68-63	6-9	24.93	8.7
S3DS11	84-63	7-9	24.17	9.1
S3DS12	114-66	11-10	22.7	8.5
S3DS13	151-66	12-13	21.89	8.6

Table 1: Description of the data sets

We assume that a document only can belongs to one cluster.

4.2. Evaluation Metric

We use the F-measure (van Rijsbergen, 1974), which is an external evaluation measure that compares the reference solution with the output of the clustering algorithm (system solution). The F-measure (F) combines the well known precision and recall measures. For a class i (reference solution) and a cluster j (system solution):

$$F(i,j) = \frac{2 \times Recall(i,j) \times Precision(i,j)}{Precision(i,j) + Recall(i,j)} \quad (4)$$

where $Recall(i,j) = \frac{n_{ij}}{n_i}$, and $Precision(i,j) = \frac{n_{ij}}{n_j}$, with n_{ij} the number of members of class i in cluster j, n_j is the number of members of cluster j, and n_i is the number of members of class i. For a clustering result, the overall F-measure is:

$$F = \sum_{i=1}^{l} \frac{n_i}{n} max_{j=1,..,k}\{F(i,j)\} \quad (5)$$

where l is the number of classes, and k is the number of clusters. F-measure values $\in [0,1]$, where the closer F-measure to 1 the value the better the clustering is. A perfect fit between the reference and the system solutions leads to a F-measure score of 1.

4.3. Results and Discussion

In Table 2 we summarize the results per collection. The first column identifies the data set. The second and third ones show the F-measure values using Binary weighting function. The sixth and seventh columns, and the tenth and eleventh ones, show the same information considering TF and TF-IDF weighting functions, respectively. The fourth, eighth and twelfth columns show the F-measure values using the NESM function. Although only one column would be enough for the NESM function, it is repeated for each weighting function to compare the results clearly. Finally, the last column shows the best F-measure result of each data set.

When we focus on the feature-weighting functions and the partial F-measure values of each data set (see boldface values), the news representation underlying the NESM function overcame Binary and TF feature-weighting functions, but not TF-IDF. Comparing cosine and correlation measures in Binary representation to NESM measure, the latter got the best result 21 times, cosine similarity got it 15 times, and correlation coefficient 11. Comparing cosine and correlation in TF representation to NESM measure, the latter got the best result 20 times, cosine similarity got it 8 times, and correlation coefficient 14. And finally, comparing cosine and correlation in TF-IDF representation to NESM measure, the latter got the best result 12 times, cosine similarity got it 20 times, and correlation coefficient 18. We present a summary of these results in Table 3.

Notice that the collection with which NESM did not achieve the best results was

S2. This collection was originally created to evaluate multilingual comparable corpora topic creation and relevant assessment, and we took the category label of the news to build the reference solution. In this case, no human revision was carried out. Thus, the data sets from this collection have more documents per cluster than the other ones, and the topics of the clusters in this collection are more general than the topics of the other collections, so that the average number of shared NEs between documents could be lower. On the other hand, NESM performs better with S1 and S2, collections with small clusters focused in one topic.

5. Conclusions and future work

We have presented a new measure, NESM, to calculate how similar two documents are. Our approach for Multilingual News Clustering is based on the representation of the news by means of the cognate Named Entities they contain. This new measure benefits from this representation and it is based only on the number and category of the NEs shared by documents.

We tested the new measure with a clustering algorithm of the CLUTO library, and we compared the obtained results with two well known similarity measures: cosine similarity and correlation coefficient. We used three collections of multilingual news to evaluate the proposed measure and we represented the news using three well known weighting functions: Binary, TF, and TF-IDF.

A proximity measure that takes into account the number and category of the NEs shared by news documents, seems to be a good way to compare multilingual news. The proposed measure NESM is competitive compared to standard similarity measures. NESM performs better than cosine and correlation measures when the news documents are represented with the Binary and TF weighting functions. NESM also performs better than the other two similarity measures when the content of the expected clusters is homogeneous, that is when they contain news of a very specific topic. When the expected clusters contain news of a very general topic, both cosine and correlation measures perform better.

On the other hand, the main advantage of using only cognate NEs for Multilingual News Clustering is that no translation resources are needed. However, the cognate identification approach requires the languages involved in the corpora to have the same alphabet and linguistic family.

The proposed measure NESM, although only computes the number of shared NEs between documents with no frequency information, overcomes standard similarities when the weighting function is TF, that considers frequency information. For this reason, we will include on NESM frequency information, with the aim to improve the obtained results with TF-IDF weighting function. In addition, we will use Okapi BM25 feature weighting, since recently this feature weighting has been seriously considered in document clustering. Finally, we will evaluate to weight different the shared NEs depending of their category.

References

Armour, Q., N. Japkowicz, and S. Matwin. 2005. The Role of Named Entities in Text Classification. In *Proceedings of CLiNE'05*.

Baeza-Yates, R. and B. Ribeiro-Neto. 1999. *Modern Information Retrieval*. ACM Press.

Carreras, X., I. Chao, L. Padró, and M. Padró. 2004. FreeLing: An Open-Source Suite of Language Analyzers . In *Proceedings of LREC04*.

Chau, R., C. Yeh, and K. Smith. 2005. A Neural Network Model for Hierarchical Multilingual Text Categorization. In *Advances in Neural Networks*, volume 3497 of *Lecture Notes in Computer Science*.

Cheung, P., R. Huang, and W. Lam. 2004. Financial Activity Mining from Online Multilingual News. In *Proceedings of the ITCC'04*.

Denicia-Carral, C., M. Montes-Gómez, L. Villaseñor-Pineda, and R. M. Aceves-Pérez. 2010. Bilingual document clustering using translation-independent features. In *Proceedings of CICLing'10*.

Dhillon, I. S. and D. S. Modha. 2001. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1):143–175.

Flaounas, I., O. Ali, M. Turchi, T. Snowsill, F. Nicart, T. De Bie, and N. Cristianini. 2011. NOAM: news outlets analysis

and monitoring system. In *Proceedings of SIGMOD*. ACM.

Gael, J. Van and X. Zhu. 2007. Correlation clustering for crosslingual link detection. In *Proceedings of IJCAI'07*.

Karypis, G. 2003. Cluto: A clustering toolkit. Technical Report 02-017, University of Minnesota, Department of Computer Science, Minneapolis.

Kogan, J., M. Teboulle, and C. Nicholas. 2005. Data Driven Similarity Measures for k-Means Like Clustering Algorithms. *Information Retrieval*, pages 331–349.

Kumaran, Giridhar and James Allan. 2004. Text classification and named entities for new event detection. In *Proceedings of SIGIR'04*. ACM.

Lawrence, J. L. 2003. Newsblaster russian-english clustering performance analysis. Technical Report CUCS-010-03, Department of Computer Science, Columbia University, New York.

Levenshtein, V. I. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707–710.

Mathieu, B., R. Besançon, and C. Fluhr. 2004. Multilingual document clusters discovery. In *Proceedings of RIAO'04*, pages 116–125.

Montalvo, S., R. Martínez, A. Casillas, and V. Fresno. 2007a. Multilingual News Document Clustering: Feature Translation vs. Identification of Cognate Named Entities. *Pattern Recognition Letters*, 28:2305–2311.

Montalvo, S., R. Martínez, A. Casillas, and V. Fresno. 2007b. Bilingual news clustering using named entities and fuzzy similarity. In *Proceedings of TSD'07*.

Pouliquen, B., R. Steinberger, C. Ignat, E. Ksper, and I. Temikova. 2004. Multilingual and cross-lingual news topic tracking. In *Proceedings of COLING '04*.

Ratinov, L. and D. Roth. 2009. Design Challenges and Misconceptions in Named Entity Recognition. In *Proc. of CoNLL '09*.

Rodríguez, H. 2002. Similitud Semantica. In *Actas del Seminario de Industrias de la Lengua de la Fundacion Duques de Soria*.

Salton, G. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

Savoy, J. 2003. Report on CLEF-2003 Multilingual Tracks. *Results of the CLEF-2003, cross-language evaluation forum*.

Shah, C., W. Bruce Croft, and D. Jensen. 2006. Representing documents with named entities for story link detection (SLD). In *Proceedings of CIKM '06*. ACM.

Shinyama, Y. and S. Sekine. 2004. Named entity discovery using comparable news articles. In *Proceedings of COLING '04*. ACL.

Silva, J., J. Mexia, C. Coelho, and G. Lopes. 2004. A Statistical Approach for Multilingual Document Clustering and Topic Extraction form Clusters. *Pliska Studia Mathematica Bulgarica*, 16:207–228.

Steinberger, R., B. Pouliquen, and J. Hagman. 2002. Cross-Lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC. In *Proceedings of CICLing '02*.

Steinberger, R., B. Pouliquen, and C. Ignat. 2005. Navigating multilingual news collections using automatically extracted information. *Journal of Computing and Information Technology*, 4:257–264.

Steinberger, R., B. Pouliquen, and C. Ignat. 2006. Exploiting multilingual nomenclatures and language-independent text features as an interlingua for cross-lingual text analysis applications. Technical Report cs.CL/0609064, EC - JRC.

Urizar, X. Saralegi and I. Alegría Loinaz. 2007. Similitud entre documentos multilingües de carácter científico-técnico en un entorno web. *Procesamiento del Lenguaje Natural*, 39:71–78.

van Rijsbergen, C. J. 1974. Foundations of evaluation. *Journal of Documentation*, 30:365–373.

Wu, K. and B. Lu. 2007. Cross-lingual document clustering. In *Proceedings of PAKDD'07*.

	COS Bin	**CORR Bin**	**NESM**	**COS TF**	**CORR TF**	**NESM**	**COS TF-IDF**	**CORR TF-IDF**	**NESM**	**Best** Global F-m.
S1DS1	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
S1DS2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
S1DS3	**0.99**	**0.99**	0.92	0.89	0.89	0.92	**0.99**	**0.99**	0.92	0.99
S1DS4	0.93	0.93	**0.94**	0.92	0.92	**0.94**	0.92	0.86	**0.94**	0.94
S1DS5	0.93	0.93	**0.95**	0.91	0.91	**0.95**	0.91	0.86	**0.95**	0.95
S1DS6	0.90	0.90	**0.96**	0.91	0.91	**0.96**	0.93	0.89	**0.96**	0.96
S1DS7	0.92	0.92	**0.95**	0.91	0.91	**0.95**	0.89	0.89	**0.95**	0.95
S1DS8	0.92	**0.94**	0.93	0.91	0.91	**0.93**	0.96	**0.97**	0.93	0.97
S1DS9	**0.94**	0.93	0.92	0.91	0.91	**0.92**	**0.97**	0.93	0.92	0.97
S1DS10	**0.93**	0.91	**0.93**	0.88	0.88	**0.93**	0.92	0.90	**0.93**	0.93
S1DS11	**0.91**	0.89	0.85	0.85	0.85	0.85	**0.90**	**0.90**	0.85	0.91
S1DS12	**0.89**	0.88	0.88	0.85	0.85	**0.88**	**0.90**	**0.90**	0.88	0.90
S1DS13	0.87	**0.88**	0.87	**0.88**	**0.88**	0.87	**0.92**	**0.92**	0.87	0.92
S2DS1	0.78	**0.89**	**0.89**	0.73	**0.89**	**0.89**	0.73	**0.89**	**0.89**	0.89
S2DS2	0.83	0.83	**0.86**	0.80	0.80	**0.86**	0.83	0.83	**0.86**	0.86
S2DS3	0.87	0.87	**0.89**	0.84	0.84	**0.89**	0.87	0.87	**0.89**	0.89
S2DS4	**0.96**	**0.96**	0.78	**0.93**	**0.93**	0.78	**0.90**	0.87	0.78	0.96
S2DS5	**0.97**	0.94	0.83	**0.94**	**0.94**	0.83	0.91	**0.94**	0.83	0.97
S2DS6	**0.96**	**0.96**	0.79	**0.94**	**0.94**	0.79	0.92	**0.94**	0.79	0.96
S2DS7	**0.95**	**0.95**	0.80	**0.95**	**0.95**	0.80	**0.95**	**0.95**	0.80	0.95
S2DS8	**0.92**	**0.92**	0.69	0.91	**0.92**	0.69	**0.93**	0.92	0.69	0.93
S2DS9	**0.92**	**0.92**	0.74	0.90	**0.92**	0.74	**0.93**	0.92	0.74	0.93
S2DS10	**0.89**	**0.89**	0.63	**0.87**	**0.87**	0.63	**0.89**	0.72	0.63	0.89
S2DS11	0.71	**0.87**	0.57	0.68	**0.70**	0.57	**0.85**	0.69	0.57	0.87
S2DS12	**0.86**	0.85	0.55	**0.63**	**0.63**	0.55	**0.75**	0.77	0.55	0.86
S2DS13	**0.69**	0.68	0.59	0.58	**0.61**	0.59	0.58	**0.73**	0.59	0.73
S3DS1	**1.0**	0.78	**1.0**	1.0	1.0	1.0	**1.0**	0.78	**1.0**	1.0
S3DS2	0.79	0.79	**1.0**	0.88	0.83	**1.0**	0.79	0.79	**1.0**	1.0
S3DS3	0.88	0.88	**0.96**	0.88	0.88	**0.96**	**0.96**	0.88	**0.96**	0.96
S3DS4	0.85	0.85	**0.87**	**0.90**	0.85	0.87	**0.90**	**0.90**	0.87	0.90
S3DS5	0.86	0.86	**0.89**	0.83	0.83	**0.89**	**0.90**	**0.90**	0.89	0.90
S3DS6	0.89	0.89	**0.94**	0.85	0.85	**0.94**	0.93	0.93	**0.94**	0.94
S3DS7	0.84	0.84	**0.89**	0.86	0.86	**0.89**	**0.93**	**0.93**	0.89	0.93
S3DS8	0.85	0.85	**0.86**	0.88	**0.91**	0.86	**0.96**	**0.96**	0.86	0.96
S3DS9	0.86	0.86	**0.88**	0.88	**0.92**	0.88	**0.96**	**0.96**	0.88	0.96
S3DS10	0.87	0.87	**0.90**	0.87	0.87	**0.90**	0.90	0.90	0.90	0.90
S3DS11	0.83	0.83	**0.89**	0.84	0.85	**0.89**	**0.91**	**0.91**	0.89	0.91
S3DS12	0.84	0.82	**0.86**	0.80	0.80	**0.86**	0.85	**0.93**	0.86	0.93
S3DS13	0.78	0.79	**0.83**	0.81	0.81	**0.83**	0.89	**0.91**	0.83	0.91

Table 2: Results for the data sets from S1, S2 and S3 collections

	COS Bin	**CORR Bin**	**NESM**	**COS TF**	**CORR TF**	**NESM**	**COS TF-IDF**	**CORR TF-IDF**	**NESM**
S1	**5**	3	**5**	1	1	**8**	5	5	5
S2	**9**	8	3	6	**11**	3	**7**	5	3
S3	1	0	**13**	1	2	**9**	**8**	**8**	4
Total	15	11	**21**	8	14	**20**	**20**	18	12

Table 3: Summary of the best partial F-measure values