

Multilingual and cross-lingual news topic tracking

Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, Emilia Käsper & Irina Temnikova

Joint Research Centre, European Commission

T.P. 267, Via E. Fermi 1

21020 Ispra (VA), Italy

<http://www.jrc.it/langtech>

Firstname.Lastname@jrc.it

Abstract

We are presenting a working system for automated news analysis that ingests an average total of 7600 news articles per day in five languages. For each language, the system detects the major news stories of the day using a group-average unsupervised agglomerative clustering process. It also tracks, for each cluster, related groups of articles published over the previous seven days, using a cosine of weighted terms. The system furthermore tracks related news *across* languages, in all language pairs involved. The cross-lingual news cluster similarity is based on a linear combination of three types of input: (a) cognates, (b) automatically detected references to geographical place names and (c) the results of a mapping process onto a multilingual classification system. A manual evaluation showed that the system produces good results.

1 Introduction

Most large organisations, companies and political parties have a department analysing the news on a daily basis. Motivations differ, but often these organisations want to know how they and their leading members are represented in the news, or they need to know whether there has been any event they ought to know about. Examples of existing news gathering and analysis systems are Informedia¹ and the *Europe Media Monitor* (Best et al. 2002). DARPA has taken an interest in the domain and launched, in 1996, the *Topic Detection and Tracking* task² (TDT) under the TIDES program. It distinguishes three major tasks: (a) segmentation of a continuous information flow (e.g. spoken news) into individual news items, (b) detection of breaking news, i.e. of a new subject that has not previously been discussed, and (c) topic tracking, i.e. the identification of related news over time. Our task is the analysis of a multilingual collection of written news articles, which means that

segmentation (task a) is of no relevance. Neither do we present here work on the detection of new topics (task b). Instead, we focus on the topic tracking task (c), and especially on the novel aspect of *cross-lingual* tracking.

The aim of our work is to provide an automatically generated overview over the major news of each day (midnight to midnight) in the languages English, German, French, Spanish and Italian. The corpus consists of news items gathered from a large number of internet news sites world-wide, and of various subscription news wires (Best et al. 2002). The texts are thus from hundreds of different sources (feeds) which often discuss the same events. Newspapers often publish the news they receive from press agencies with no or few amendments. The corpus of news articles thus contains not only summaries of the same events written by different journalists, but also many duplicates and near duplicates of the same original text which need to be eliminated from the collection.

In order to identify the major news, we identify clusters of similar news items, i.e. news items that deal with the same subject. All subjects that trigger a large number of news articles from various feeds are of interest. The related news thus do not necessarily have to discuss *events*, i.e. things that happen at a particular time and place (e.g. the 11/03 Madrid bombing), but they can also be a thread of discussions on the same subject, such as the campaign for the US presidential elections.

In section 2, we summarise other work on topic tracking, on cross-lingual news linking and on feature extraction methods. Section 3 describes the multilingual news corpus and the text feature extraction used for the document representation. In section 4, we present the process and evaluation of major news identification. Section 5 is dedicated to the multi-monolingual topic tracking process and its evaluation. Section 6 describes the cross-lingual linking of related clusters of major news, plus evaluation results. Section 7 points to future work.

¹ <http://www.informedia.cs.cmu.edu/>

² <http://www.nist.gov/speech/tests/tdt/>

2 Related work

Allan et al. (1998) identify new events and then track the topic like in an information filtering task by querying new documents against the profile of the newly detected topic. Topics are represented as a vector of stemmed words and their TF.IDF values, only considering nouns, verbs, adjectives and numbers. In their experiments, using between 10 and 20 features produced optimal results. Schultz (1999) took the alternative approach of clustering texts with a single-linkage unsupervised agglomerative clustering method, using cosine similarity and TF.IDF for term weighting. He concludes that “a successful clustering algorithm must incorporate a representation for a cluster itself as group average clustering does”. We followed Schultz’ advice. Unlike Schultz, however, we use the log-likelihood test for term weighting as this measure seems to be better when dealing with varying text sizes (Kilgarriff 1996). We do not consider parts-of-speech, lemmatisation or stemming, as we do not have access to linguistic resources for all the languages we need to work with, but we use an extensive list of stop words.

Approaches to *cross-lingual* topic tracking are rather limited. Possible solutions for this task are to either translate documents or words from one language into the other, or to map the documents in both languages onto some multilingual reference system such as a thesaurus. Wactlar (1999) used bilingual dictionaries to translate Serbo-Croatian words and phrases into English and using the translations as a query on the English texts to find similar texts. In TDT-3, only four systems tried to establish links between documents written in different languages. All of them tried to link English and Chinese-Mandarin news articles by using Machine Translation (e.g. Leek et al. 1999). Using a machine translation tool before carrying out the topic tracking resulted in a 50% performance loss, compared to monolingual topic tracking.

Friburger & Maurel (2002) showed that the identification and usage of proper names, and especially of geographical references, significantly improves document similarity calculation and clustering. Hyland et al. (1999) clustered news and detected topics exploiting the unique combinations of various named entities to link related documents. However, according to Friburger & Maurel (2002), the usage of named entities alone is not sufficient.

Our own approach to cross-lingual topic tracking, presented in section 6, is therefore based on three kinds of information. Two of them exploit the co-occurrence of named entities in related news stories: (a) cognates (i.e. words that are the same across languages, including names) and (b) geographical references. The third component, (c) a

process mapping texts onto a multilingual classification scheme, provides an additional, more content-oriented similarity measure. Pouliquen et al. (2003) showed that mapping texts onto a multilingual classification system can be very successful for the task of identifying document translations. This approach should thus also be an appropriate measure to identify *similar* documents in other languages, such as news discussing the same topic.

3 Feature extraction for document representation

The similarity measure for monolingual news item clustering, discussed in section 4, is a cosine of weighted terms (see 3.1) enriched with information about references to geographical place names (see 3.2). Related news are tracked over time by calculating the cosine of their cluster representations, while setting certain thresholds (section 5). The cross-lingual linking of related clusters, as described in section 6, additionally uses the results of a mapping process onto a multilingual classification scheme (see 3.3).

The news corpus consists of a daily average of 3350 English news items, 2100 German, 870 Italian, 800 French and 530 Spanish articles, coming from over three hundred different internet sources.

3.1 Keyword identification

For monolingual applications, we represent documents by a weighted list of their terms. For the weighting, we use the log-likelihood test, which is said to perform better than the alternatives TF.IDF or chi-square when comparing documents of different sizes (Kilgarriff 1996). The reference corpus was produced with documents of the same type, i.e. news articles. It is planned to update the reference word frequency list daily or weekly so as to take account of the temporary news bias towards specific subjects (e.g. the Iraq war). We set the p-value to 0.01 in order to limit the size of the vector to the most important words. Furthermore, we use a large list of stop words that includes not only function words, but also many other words that are not useful to represent the contents of a document. We do not consider part-of-speech information and do not carry out stemming or lemmatisation, in order to increase the speed of the process and to be able to include new languages quickly even if we do not have linguistic resources for them. Clustering results do not seem to suffer from this lack of linguistic normalisation, but when we extend the system to more highly inflected languages, we will have to see whether lemmatisation will be necessary. The result of the keyword identification process is thus a representation of each incoming news article in a vector space.

3.2 Geographical Place Name Recognition

For place name recognition, we use a system that has been developed by Pouliquen et al. (2004). Compared to other named entity recognition systems, this tool has the advantage that it recognises exonyms (foreign language equivalences, e.g. *Venice* vs. *Venezia*) and that it disambiguates between places with the same name (e.g. *Paris* in France vs. the other 13 places called *Paris* in the world). However, instead of using the city and region names as they are mentioned in the article, each place name simply adds to the *country score* of each article. The idea behind this is that the place names themselves are already contained in the list of keywords. By adding the country score separately, we heighten the impact of the geographical information on the clustering process.

The country scores are calculated as follows: for each geographical place name identified for a given country, we add one to the country counter. We then normalise this value using the log-likelihood value, using the average country counter in a large number of other news articles as a reference base. As with keywords, we plan to update the country counter reference frequency list on a daily or weekly basis. The resulting normalised country score has the same format as the keyword list so that it can simply be added to the document vector space representation.

3.3 Mapping documents onto a multilingual classification scheme

For the semantic mapping of news articles, we use an existing system developed by Pouliquen et al. (2003), which maps documents onto a multilingual thesaurus called *Eurovoc*. Eurovoc is a wide-coverage classification scheme with approximately 6000 hierarchically organised classes. Each of the classes has exactly one translation in the currently 22 languages for which it exists. The system carries out category-ranking classification using Machine Learning methods. In an inductive process, it builds a profile-based classifier by observing the manual classification on a training set of documents with only positive examples. The outcome of the mapping process is a ranked list of the 100 most pertinent Eurovoc classes. Due to the multilingual nature of Eurovoc, this representation is independent of the text language so that it is very suitable for cross-lingual document similarity calculation, as was shown by Pouliquen et al. (2003).

4 Clustering of news articles

In this process, larger groups of similar articles are grouped into clusters. Unlike in document classification, clustering is a bottom-up, unsupervised

process, because the document classes are not known beforehand.

4.1 Building a dendrogram

In the process, we build a hierarchical clustering tree (dendrogram), using an agglomerative algorithm (Jain et al. 1999). In a first step, (1) we calculate the similarity between each document pair in the collection (i.e. one full day of news in one language), applying the cosine formula to the document vector pairs. The vector for each single document consists of its keywords and their log-likelihood values, enhanced with the country profile as described in sections 3.1 and 3.2. (2) When two or more documents have a cosine similarity of 90% or more, we eliminate all but one of them as we assume that they are duplicates or near-duplicates, i.e. they are exact copies or slightly amended versions of the same news wire. (3) We then combine the two most similar documents into a cluster, for which we calculate a new representation by merging the two vectors into one. For the node combining the two documents, we also have an intra-cluster similarity value showing the degree to which the two documents are similar. For the rest of the clustering process, this node will be treated like a single document, with the exception that it will have twice the weight of a single document when being merged with another document or cluster of documents. We iteratively repeat steps (1) and (3) so as to include more and more documents into the binary dendrogram until all documents are included. The resulting dendrogram will have clusters of articles that are similar, and a list of keywords and their weight for each cluster. The degree of similarity for each cluster is shown by its intra-cluster similarity value.

4.2 Cluster extraction to identify main events

In a next step, we search the dendrogram for the major news clusters of the day, by identifying all sub-clusters of documents that fulfil the following conditions: (a) the intra-cluster similarity (cluster cohesiveness) is above the threshold of 50%; (b) the number X of articles in the cluster is at least 0.6% of the total number of articles of that language per day; (c) the number Y of different feeds is at least half the minimum number of articles per cluster ($Y = X/2$).

The threshold of 50% in (a) was chosen because it guarantees that most related articles are included in the cluster, while unrelated ones are mostly excluded (see section 4.3). The minimum number of articles per cluster in (b) was chosen to limit the number of major news clusters per day. We requested a minimum number of different news feeds (c) so as to be sure that the news items are of

general interest and that we are not dealing with some newspaper-specific or local issues.

With the current settings, the system produces an average of 9 English major news clusters per day, 11 Italian, 16 German, 20 French and 21 Spanish. The varying numbers indicate that the settings should probably be changed so as to produce a similar number of major news clusters per day in the various languages. Most likely, the minimum number of feeds should have an upper maximum value for languages like English with thousands of news articles per day.

For each cluster, we have the following information: number of articles, number of sources (feeds), intra-cluster similarity measure and keywords. Using our group-average approach we also have the centroid of the cluster (i.e. the vector of features that represents the cluster). For each cluster, we compute the article that is most similar to the centroid (short: the *centroid article*). We use the title of this centroid article as the title for the cluster and we present this article to the users as a first document to read about the contents of the whole cluster.

The collection of clusters is mainly presented to the users as a flat and independent list of clusters. However, as we realised that some of the clusters are more related than others (e.g. with the recent interest in Iraq, there are often various clusters covering different aspects of the political situation of the country), we position clusters with an inter-cluster similarity of over 30% closer to each other when presenting them to the users.

4.3 Evaluation of the monolingual clustering

The evaluation of clustering results is rather tricky. According to Joachims (2003), clustering results can be evaluated using a variety of different ways: (a) let the market decide (select the winner); (b) ask end users; (c) measure the ‘tightness’ or ‘purity’ of clusters; (d) use human-identified clusters to evaluate system-generated ones. The last solution (d) is out of our reach because it is very resource-consuming; several evaluators would be needed for cross-checking the human judgement. The ‘market’ (a) and user groups (b) will use and evaluate our system in the near future, but we need to evaluate the system prior to showing it to a large number of customers. We therefore focus on method (c) by letting a person judge how consistently the articles of each cluster treat the same story.

We evaluated the major clusters of English news articles (using the 50% intra-cluster similarity threshold) produced for the seven-day period starting 9 March 2004. During this period, 71 clusters containing 1072 news articles were produced. The

evaluator was asked to decide, for each cluster and on a four-grade scale, to what extent the clustered articles were related to the centroid article. Comparing the clustered articles to the centroid article was chosen over evaluating the homogeneity of the cluster because it is both easier and closer to the real-life situation of the users: users will enter the cluster via the centroid article and will judge the other articles according to whether or not they contain the information they expect. The evaluation scale distinguishes the following ratings:

- (0) *wrong link*, e.g. Madrid football results vs. Madrid elections; this is a hypothetical example as no such link was found.
- (1) *loosely connected* story, e.g. Welsh documentary on drinking vs. alcohol policy in Britain;
- (2) *interlinked news stories*, e.g. 11/03 Madrid bombing vs. elections of the Spanish Prime Minister Zapatero vs. Spanish decision to pull troops out of Iraq;
- (3) *same news story*.

In the evaluation, 91.5% of the articles were rated as good (3), 7.7% were rated as interlinked (2) and 0.8% were rated as loosely connected. No wrong links were found. 47 of the 71 clusters only contained good articles (3). Loosely connected articles (1) were distributed evenly. No more than two articles of this rating were found in a single cluster. They never amounted to more than 17% of all articles in a cluster (2 out of 12 articles).

An evaluation of the clusters produced on one day’s data with 30% and 40% intra-cluster similarity thresholds showed that the performance decreased drastically. In 30%-clusters, we found several wrong links (category 0), while no such wrong links were found in the 50%-clusters. The total number of wrong (0) or loosely connected (1) articles went up from one (in the 50%-cluster for that day) to 37. Furthermore, the worst clusters contained over 50% of such unrelated articles. The 40%-clusters were of a slightly better quality, but they still were clearly less good than the 50%-clusters: The percentage of wrong (0) and loosely connected (1) articles only went up from 0.8% (in the 50%-clusters) to 4%, but some of the 40%-clusters still had more bad (category 0 or 1) than good (category 2 or 3) articles. These numbers confirm that our choice of the 50% intra-cluster similarity threshold is most useful.

We have not produced a quantitative evaluation of the miss rate of the clustering process (i.e. the number of related articles *not* included in the cluster, showing the *recall*). However, a full-text search of the relevant proper names in the rest of the news collection showed that the clustering

process missed very few related articles. In any case, from our users' point of view, it is much more important to know the major news stories of a specific day than being able to access all articles on the subject.

Statistical evaluation showed no correlation between cluster size and accuracy. However, category (2) results were more frequently found in clusters pertaining to news stories that go on for a long time, such as the US presidential elections. These stories get wide coverage without being 'breaking news', and many of the articles involved are commentaries. Some of the category (2) results were also found in stories around the Madrid bombing and its consequences: some articles discussed the bombing itself on 11 March (number of dead, investigation, mourning); others discussed the fact that, in the 14 March elections, the Spanish people elected the socialists as they felt that former Prime Minister Aznar's politics were partially responsible for this tragedy; yet other articles discussed the post-election consequences such as the decision of the new Socialist government to pull out the Spanish troops from Iraq, etc. Many of the articles touched upon several of these issues. Articles were rated as good (3) if they had at least one core topic in common with the centroid article.

5 Monolingual linking of news over time

Establishing automatic links between the major clusters of news published in one language in the last 24 hours and the news published in previous days can help users in their analysis of events. Establishing historical links between related news stories is the third of the TDT tasks (see the introduction in section 1).

We track topics by calculating the cosine similarity between all major news clusters of one day with all major news clusters of the previous days, currently up to a maximum distance of seven days. The input for the similarity calculation is the cluster vector produced by the monolingual clustering process (see section 4.2). The output for each pairwise similarity calculation is a similarity value between 0 and 1. Whether we decide that two clusters are related or not depends on the similarity threshold we set. We found that related clusters over time have an extremely high similarity, often around 90%, which shows that the vocabulary used in news stories over time changes very little. For testing purposes, we set the threshold very low, at 15%, so that we could determine a useful threshold during the evaluation process.

5.1 Evaluation of historical linking

We evaluated the historical links for the 136 English clusters of major news produced for the

two-week period starting on 9 March 2004, looking at the seven-day window preceding the day for which each major news cluster was identified. The total number of historical links found for this period is 228, i.e. on average 1.68 historical links per major news cluster. However, for 42 of the 136 major news clusters, the system did not find any related news clusters with a similarity of 15% or more.

We made a binary distinction between 'closely related articles' (+) and 'unrelated, or not so related articles' (-). The evaluation results at varying cosine similarity thresholds, displayed in Table 1, show that there is no threshold which includes all good clusters and excludes all bad ones. Setting the threshold at 40% would mean that 173 (135+24+14) of the 203 good clusters (86%) would be found while three bad ones would also be shown to the user. Setting the threshold at the more inclusive level of 20% would mean that 199 of the 203 good clusters (98%) would be found, but the number of unrelated ones would increase to 17.

Similarity	+ Related	- Unrelated
15 – 19%	4	8
20 – 39%	26	14
40 – 59%	14	2
60 – 79%	24	0
80 – 100%	135	1
Total	203	25

Table 1: Evaluation, for varying similarity thresholds, of the automatically detected links between major news of the day and the major news published in the seven days before. The distinction was binary: *Related* (+) or *Not (so) related* (-).

6 Cross-lingual linking of news clusters

News analysts and employees in press rooms and public relations departments often want to see how the same news is discussed in different countries. To allow easy access to related news in other languages, we establish cross-lingual links between the clusters of major news stories. As major news in one country sometimes is only minor news in another, we calculate a second, alternative group of news clusters for each language and each day, containing a larger number of smaller clusters. To get this alternative group of clusters, we set the intra-cluster similarity to 25% and require that the news of the cluster come from at least two different news sources. These conditions are much weaker than the requirements described in section 4.2. For each major news cluster (50% intra-cluster similar-

ity) per day and per language, we thus try to find related news in the other languages among any of the smaller clusters produced with the 25% intra-cluster similarity requirement.

We use three types of input for the calculation of cross-lingual cluster similarity: (a) the vector of keywords, as described in section 3.1, not enhanced with geographical information, (b) the country score vector, as described in section 3.2, and (c) the vector of Eurovoc descriptors, as described in section 3.3. The impact of the three components is currently set to 20%, 30% and 50% respectively. Using the Eurovoc vector alone would give very high similarity values for, say, news about elections in France and in the United States. By adding the country score, a considerable weight in the cross-lingual similarity calculation is given to the countries that are mentioned in each news cluster. The overlap between the keyword vectors of documents in two different languages will, of course, be extremely little, but it increases with the number of named entities that the documents have in common. According to Gey (2000), 30% of content-bearing words in journalistic text are proper names.

The system ignores individual articles, but calculates the similarity between *whole clusters* of the different languages. The country score and the Eurovoc descriptor vector are thus assigned to the cluster as a whole, treating all articles of each cluster like one big bag of words.

6.1 Evaluation of cross-lingual cluster links

The evaluation for the cross-lingual linking was carried out on the same corpus as the evaluation of the historical links, i.e. taking the 136 English major news clusters as a starting point. Cross-lingual cluster links were evaluated for two languages, English to French and English to Italian. The evaluation was again binary, i.e. clusters were either judged as being ‘closely related’ (+) or ‘unrelated, or not so related’ (−). For 31 English clusters, no French cluster was found. Similarly, for 32 English clusters, no Italian cluster was found. This means that for almost 25% of the English-speaking major news stories (31/136), there was no equivalent news cluster in the other languages.

For the remaining English clusters, a total of 131 French and 133 Italian clusters were detected by the system, i.e. on average more than one for each English cluster. However, when several related news clusters were found, only the one with the highest score was considered in the evaluation.

Table 2 not only shows that the English-Italian links are less reliable than the English-French ones (the Italian document representation is inferior to the French one because we spent less effort on op-

timising the Italian keyword assignment), but also that the quality of cross-lingual links is generally lower than the historical links presented in section 5.1. If we set the threshold for identifying related news across languages to 30%, the system catches 74 of the 75 good French clusters (99%) and 67 of the 69 Italian clusters (97%). However, the system then also proposes 13 bad French and 12 bad Italian clusters to the users. Setting the threshold higher would decrease the number of wrong hits. However, we decided to use the threshold of 30% because we consider it important for users to be able to find related news in other languages. Furthermore, unrelated clusters are usually very easy to detect just by looking at the title of the cluster.

Similarity	FR +	FR −	IT +	IT −
15 – 19%	0	7	0	1
20 – 29%	1	6	2	11
30 – 39%	5	6	7	8
40 – 49%	16	4	13	5
50 – 59%	19	1	18	6
60 – 100%	34	1	29	1
Total	75	25	69	32

Table 2: Evaluation, for varying similarity thresholds, of the automatically detected cross-lingual links between English major news and French (FR) or Italian (IT) news of the same day. The distinction was binary: *Related* (+) or *Not (so) related* (−).

7 Conclusion and future work

We have shown that our system can rather accurately identify clusters of major news per day in five languages and that it can link these clusters to related news over time (topic tracking). The most interesting and novel feature of the system is, however, that it can also identify related news *across* languages, without translating articles or using bilingual dictionaries. This cross-lingual cluster similarity is achieved by a combination of three feature sets, which currently have an impact of 50%, 30% and 20%, respectively: the main feature set is the mapping onto the multilingual classification scheme Eurovoc; the others are the countries referred to in the articles (direct mention of the country, or of a smaller place name of that country) and the cognates (same strings used in the articles across languages, i.e. mainly named entities). The evaluation has shown that the results are good, but that the cross-lingual linking performs less well than the monolingual historical linking of related news clusters. Users felt that the system performs

well enough for it to go online soon, for usage by a large user community of several thousand people. Improvements to the system will nevertheless be sought.

Future work will include testing different settings concerning the relative impact of the three components, as well as detecting and using more named entities such as absolute and relative date expressions, proper names, etc. A further aim is to extend the system to another six languages.

The usage of cognate similarity could be improved. Currently it will not work with Greek, for instance, except for a few proper names. We would therefore like to experiment with multi-lingual stemming methods to exploit the existence of similar words across languages such as English *elephant*, French *éléphant*, Spanish and Italian *elefante* and German *Elefant*.

Several customer groups requested an advanced news analysis that distinguishes between articles about concrete events and articles commenting about these events. We will explore this issue, but it is very likely that this distinction will require a syntactic analysis of the news and cannot be made with our bag-of-words approach.

Finally, we intend to work on breaking news detection, i.e. detecting *new* events, as opposed to detecting major news. This work will require working on smaller time windows than the current 24-hour window.

8 Acknowledgements

We would like to thank the Web Technology group of the Joint Research Centre for their collaboration and for giving us access to their valuable multilingual news collection. Our special thanks goes to Clive Best, Erik van der Goot, Ken Blackler and Teofilo Garcia. We would also like to thank our former colleague Johan Hagman for introducing us to the methods and usefulness of cluster analysis.

References

- Allan James, Ron Papka & Victor Lavrenko (1998). *On-line New Event Detection and Tracking*. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 37-45. Melbourne, Australia
- Best Clive, Erik van der Goot, Monica de Paola, Teofilo Garcia & David Horby (2002). *Europe Media Monitor – EMM*. JRC Technical Note No. I.02.88. Ispra, Italy.
- Friburger N. & D. Maurel (2002). *Textual Similarity Based on Proper Names*. Proceedings of the workshop Mathematical/Formal Methods in Information Retrieval (MFIR'2002) at the 25th ACM SIGIR Conference, pp. 155-167. Tampere, Finland.
- Gey Frederic (2000). *Research to Improve Cross-Language Retrieval – Position Paper for CLEF*. In C. Peters (ed.): *Cross-Language Information Retrieval and Evaluation*, Workshop of Cross-Language Evaluation Forum (CLEF'2000), Lisbon, Portugal. Lecture Notes in Computer Science 2069, Springer.
- Hyland R., C. Clifton & R. Holland (1999). *Geo-NODE: Visualizing News in Geospatial Context*. In Afca99.
- Jain A., M. Murty & P. Flynn (1999). *Data clustering: a review*. Pages 264
- Joachims Thorsten (2003). *Representing and Accessing Digital Information*. Available at <http://www.cs.cornell.edu/Courses/cs630/2003fa/lectures/tclust.pdf>
- Kilgarriff A. (1996) *Which words are particularly characteristic of a text? A survey of statistical approaches*. Proceedings of the AISB Workshop on Language Engineering for Document Analysis and Recognition. Sussex, 04/1996, pp. 33-40.
- Leek Tim, Hubert Jin, Sreenivasa Sista & Richard Schwartz (1999). *The BBN Crosslingual Topic Detection and Tracking System*. In 1999 TDT Evaluation System Summary Papers. <http://www.nist.gov/speech/tests/tdt/tdt99/papers>
- Pouliquen Bruno, Ralf Steinberger & Camelia Ignat (2003). *Automatic identification of document translations in large multilingual document collections*. Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP'2003), pp. 401-408. Borovets, Bulgaria, 10 - 12 September 2003.
- Pouliquen Bruno, Ralf Steinberger, Camelia Ignat & Tom de Groeve (2004). *Geographical Information Recognition and Visualisation in Texts Written in Various Languages*. Proceedings of the 2004 ACM Symposium on Applied Computing, Session on *Information Access and Retrieval* (Nicosia, Cyprus), Volume 2 of 2, pages 1051-1058. New York.
- Schultz J. Michael & Mark Liberman (1999). *Topic detection and Tracking using idf-weighted Cosine Coefficient*. DARPA Broadcast News Workshop Proceedings.
- Wactlar H.D. (1999). *New Directions in Video Information Extraction and Summarization*. In Proceedings of the 10th DELOS Workshop, Satorini, Greece, 24-25 June 1999.