JERARTOP: A New Topic Detection System*

Aurora Pons-Porrata¹, Rafael Berlanga-Llavori²,
José Ruiz-Shulcloper³, and Juan Manuel Pérez-Martínez²

¹ Universidad de Oriente, Santiago de Cuba, Cuba
aurora@app.uo.edu.cu

² Universitat Jaume I, Castellón, Spain
{berlanga,martinej}@lsi.uji.es

³ Advanced Technologies Application Center, MINBAS, Cuba
jshulcloper@cenatav.co.cu

Abstract. In this paper we present an on-line detection system, named *JERARTOP*, which goes beyond traditional detection systems, because it generates the implicit knowledge of a stream of documents. This knowledge is expressed as a taxonomy of topics/events, which is automatically built by the system in an incremental way. Moreover, the proposed detection system also annotates each detected topic using a set of predefined subjects, as well as it provides a summary for the topic. The experimental results demonstrate its usefulness and its effectiveness as a detection system.

1 Introduction

Topic Detection and Tracking (TDT) is a new line of research that comprises three major sub-problems: segmenting speech-recognized TV/radio broadcasts into news stories, detecting novel events, and tracking the development of an event according to a given set of sample stories of that event [1]. Starting from a continuous stream of newspaper articles, the *Event Detection* problem consists in determining for each incoming document, whether it reports on a new event, or it belongs to some previously detected event.

One of the most important issues in this problem is to define what an *event* is. Initially, an event can be defined as something that happens at a particular place and time. However, many events occur along several places and several time periods (e.g. the whole event related to a complex trial). For this reason, researchers in this field prefer the broader term of *Topic*, which is defined as an important event or activity along with all its directly related events [2]. In this context, a *Topic Detection System* aims at discovering the topics reported in a stream of newspaper articles. For this purpose, incremental unsupervised clustering algorithms have been widely applied, such as the *K-Means*, *Single-Pass*, *1-NN* and others [3, 4, 5, 6].

In this paper we address three main problems that currently event detection systems present, namely: 1) they make irrevocable clustering assignments and therefore the set of detected topics can depend on the arrival order of the documents, 2) they do not take into account the different abstraction levels at which events and topics can be represented, and 3) they do not provide any summary of the detected topics, so that users can know quickly the cluster's contents.

* This work has been partially funded by the research project CICYT (TIC 2002-04586-C04-03).

The proposed solution consists of a new on-line detection system named *JERARTOP*. It generates incrementally the implicit knowledge of the incoming documents. This knowledge is expressed as a hierarchy of topics, that is, our system identifies not only the topics but also the structure of events they comprise. Moreover, the proposed system also categorizes each hierarchy topic with respect to a set of predefined subjects, and it provides a summary for it.

The remainder of the paper is organized as follows: Section 2 presents the global architecture of our system. Section 3 describes the representation of documents taking into account their temporal and place components, a similarity measure for these documents and the clustering algorithms. The description methods are presented in Section 4 and our experiments are showed in Section 5. Finally, conclusions are presented in Section 6.

2 Global Architecture

The global architecture of our detection system is presented in Figure 1. Starting from a continuous stream of newspaper articles, the system first processes them to get a proper representation to cluster them according to their semantics. The steps included in this pre-processing are the following ones: document texts are tokenised, abbreviations, proper names and some simple phrases are recognised, stop-words are removed, part-of-speech tags are added and common words are replaced by their lemmas. All these elements are called terms and they will form a vector with their text frequency. Additionally, the system also extracts some useful metadata such as the publication dates and the places mentioned in the texts. Thus, the document representation can take into account the time and place properties.

Fig. 1. Global architecture of the detection system *JERARTOP*.

3 Document Clustering Algorithm

Each incoming document d^i is represented in our system with three components:

- A vector of weighted terms $T^i = (TF_1^i, \dots, TF_n^i)$, where the terms represent the lemmas of the words appearing in the content of the document, and TF_k^i is the relative frequency of the term t_k in d^i .

- *A vector of publication dates.* This vector is only formed by one publication date when comparing documents instead of cluster representatives.
- *A vector of weighted places* $(TF_{p_1^i}, \dots, TF_{p_{l_i}^i})$, where $TF_{p_k^i}$ is the absolute frequency of the place p_k in d^i and $k=1, \dots, l_i$. Place names are automatically extracted from the content of the documents by using a thesaurus. They are represented by their paths within the thesaurus, where each level indicates a geographic region.

Automatic document clustering, as in event detection, is based on a similarity measure and a clustering criterion. The cosine measure mostly uses to compare two documents. In our case, we consider that two documents refer to the same event if their contents, places and time references approximately coincide.

To compare the term vectors of two documents d^i and d^j we use the cosine measure: $S_T(d^i, d^j) = \cos(T^i, T^j)$.

To compare the time vectors of two documents we propose the following function:

$$D(d^i, d^j) = \min_{f^i \in FR^i, f^j \in FR^j} \{d(f^i, f^j)\}, \text{ where } d(f^i, f^j) \text{ is the number of days between both and } FR^i \text{ is the set of all publication dates } f^i \text{ of the document } d^i.$$

To compare the place vectors of two documents we propose the following Boolean function:

$$S_P(d^i, d^j) = \begin{cases} 1 & \text{if } \exists p_q^i, p_l^j \text{ such that they have a common prefix} \\ 0 & \text{otherwise} \end{cases}$$

Finally, the global similarity measure can be defined as follows:

$$S(d^i, d^j) = \begin{cases} S_T(d^i, d^j) & \text{if } S_P(d^i, d^j) = 1 \wedge D(d^i, d^j) \leq \beta_{time} \\ 0 & \text{otherwise} \end{cases}$$

where β_{time} is the maximum number of days that are required to determine whether two articles refer to the same or to different events. This measure tries to capture the idea that two documents reporting a same event should have a high semantic similarity, time proximity and place coincidence.

Hierarchical Document Clustering

One of the main novelties of *JERARTOP* is that it applies an incremental hierarchical clustering algorithm to the detection problem. This algorithm attempts to build a hierarchy of topics by clustering successively the representatives of the clusters obtained in the previous hierarchy level. In this way, the higher is the hierarchy level we have less details in the detected topics. The lowest level of the hierarchy is intended to allocate the smallest detectable events, which are groups of tightly related documents. The rest of the levels contain cluster representatives. The representative of a cluster c , denoted as $\bar{c}$, is also a document, represented as $(T^{\bar{c}}, F^{\bar{c}}, P^{\bar{c}})$, which is calculated as the union of the documents in c .

JERARTOP uses as the basic clustering routine the compact clustering algorithm presented in [7], which is based on the incremental construction of β_0 -compact sets. In

this way the system ensures that the detected sets are independent of the document arrival order.

In order to ensure this property in the hierarchy, the system must revise all the detected topics in the hierarchy every time a new document arrives. For this purpose, the system first applies the clustering routine in the lowest level of the hierarchy. New clusters could appear and other existing could disappear. When clusters are removed from any level of the hierarchy, their representatives must be removed from the clusters in which they were located in the next level. Similarly, when new clusters are created, their representatives must be calculated. The members of the clusters where changes took place (that is, some cluster representatives were eliminated) as well as all new representatives, must be queued in order to incorporate them to the set of existing clusters in this level of the hierarchy. For that purpose, we apply the clustering routine again. This process is repeated until the top level on the hierarchy is reached [8].

4 Cluster Description Methods

In order to determine at a glance whether the content of an event are of user interest, we propose a new summarization method based on Testor Theory [9]. Starting from a set of document clusters, each one representing a different event or topic, our method selects the frequent terms of each cluster that are not included in the other clusters. Once these terms have been selected, the system extracts all the sentences that contain the selected terms. Finally, in order to improve the coherence and organization of the summaries, we sort the extracted sentences according to the publication date of the news and their position in the text.

For each cluster c , we construct a matrix $MR(c)$, whose columns are the most frequent terms in the representative $\bar{c}$ and its rows are the representatives of all clusters, described in terms of these columns. In order to calculate the typical testors, we considered two classes in the matrix $MR(c)$. The first class is only formed by $\bar{c}$ and the second one is formed by all remaining cluster representatives. The comparison criterion applied to all the features is:

$$d(v_{i_k}, v_{j_k}) = \begin{cases} 1 & \text{if } v_{i_k} - v_{j_k} \geq \delta \\ 0 & \text{otherwise} \end{cases},$$

where v_{i_k}, v_{j_k} are the frequencies in the cluster representative i and j in the column corresponding to the term t_k respectively, and δ is an user-defined parameter.

Then, a summary of an event c consists of a set of sentences extracted from the documents in c , in which the highest quantity of terms that belong to the maximum length typical testors of the matrix $MR(c)$ occurs. Moreover, the sentences that cover the calculated typical testor set are also added to the summary.

In addition to this, *JERARTOP* also assigns subject matters to each detected topic of the hierarchy. For this purpose, we have adopted the subjects proposed by the International Press Telecommunications Council (IPTC)¹. The IPTC Subject Reference System provides a list of internationally acceptable subject names, subject matter names and some subject detail names for categorizing the content of news.

¹ <http://www.iptc.org/>

To assign IPTC subjects to detected topics, we make use of the synsets of the lexical database *WordNet*². We have used the different relationships between synsets in order to annotate each synset of *WordNet* with an IPTC code. More specifically, we first manually associate to each IPTC subject a proper synset. Then the system automatically propagates this IPTC code to all the directly related synsets and its hyponyms. Notice that not all the synsets will have associate an IPTC code.

Finally, in order to assign an IPTC subject to each detected topic, the system applies the following method. Given the cluster representative of the topic, the system first obtains a tuple with the synsets associated to the term vector. Notice that these terms can have several senses, and therefore some disambiguation method must be applied to reduce the number of ambiguous synsets. In this work we have used the method presented in [10]. Afterwards, the system determines the 10 IPTC subjects that annotate more synsets in the disambiguated representative. As the IPTC subjects are organized into taxonomy, each selected IPTC subject votes for its ancestors in this taxonomy. The vote is the number of the synsets annotated with this subject in the representative. Finally, the IPTC subject that has the maximum number of votes is selected. The final subject of the cluster will be the IPTC that satisfies the following conditions:

- It is one of the 10 IPTC subjects initially selected.
- It is a descendent of the IPTC subject that received the maximum number of votes.
- It annotates the greater number of the synsets in the cluster representative among all descendents of the IPTC subject that received the maximum number of votes.

The intuitive idea of this method is to select firstly, the more general IPTC subject that annotates the greater number of terms in the document cluster and then, inside this subject matter, to select the most particular subject.

5 Experiments and Results

The effectiveness of our detection system has been evaluated using two collections. The first one contains 554 articles published in the Spanish newspaper "El País" during June 1999. We have manually constructed a hierarchy of topics. In this work we only deal with the two first levels of this hierarchy, called the *Event Level* and *Topic Level* respectively. In the first level we have identified 85 non-unitary events. From these events we have identified 58 topics. The collection covers 21 events associated to the end of the "Kosovo War" along with their immediate consequences, the visit of the Pope to Poland, the elections in several countries, among others.

To evaluate the results we use two measures of the literature that compare the system-generated clusters with the manually labeled topics, namely: the F1-measure and the Detection Cost [2]. Table 1 shows the minimum topic-weighted cost and the maximum F1-Measure obtained by our system on the "El País" dataset in both levels of the hierarchy. The first line shows results for the base system using the cosine metric. Line 2 shows results when using our similarity measure instead of cosine distance. Our similarity measure performs much better with respect the F1-Measure, improving the results by 0.06 in both levels. As a consequence, we can conclude that the time and spatial components improves notably the quality of the system-generated topics. These components do not affect to the Detection Cost.

² <http://www.cogsci.princeton.edu/~wn/>

Table 1. The obtained results in “El País” collection.

System	Event Level		Topic Level	
	Detection Cost	F1-Measure	Detection Cost	F1-Measure
Base cosine	0.0027	0.7237	0.0039	0.7572
<i>JERARTOP</i>	0.0026	0.7848	0.0039	0.8114

Figure 2 shows a snippet of the created hierarchy by *JERARTOP*. This hierarchy is represented as a XML file containing for each topic the events that it comprises and its subject matter. Each event contains also the titles of the news, its subject matter and its summary. The carried out experiments demonstrate the usefulness of the description methods. The summaries capture the main ideas about each topic and an appreciable reduction of words is also achieved (superior to 85% in most of the topics). In case of the subject matters, we evaluate the results obtained in each topic in good (G), bad (B) and not too bad (R). Table 2 shows for some topics of this collection, the obtained subject and its evaluation. Considering alone the good answers, the method obtains a precision of 71%, whereas if we also consider as good those classified as not too bad, we obtain a precision of 88%.

```

- <Topic Id= "Cluster_134" Level= "1">
  <Subject IPTC= "Politit & Treaty and International Organisations & alliances" />
  - <Event Id= "Cluster_166" Level= "1">
    <Description Summary= "Ni siquiera una conversación telefónica celebrada la semana pasada entre Chirac, Aznar y el presidente argentino, Carlos Menem, ha modificado la decisión francesa de retrasar la apertura de negociaciones con Chile y con el Mercosur. El viceministro de Exteriores alemán, Günther Verheugen, expresó el deseo de la presidencia alemana de aprobar los mandatos de negociación antes de esa cita que reunirá a los jefes de Estado o de Gobierno de todos los países de América Latina y del Caribe y de la UE. Los países de Mercosur y Chile advirtieron a la UE el pasado fin de semana en México que quieren un acuerdo de libre comercio y que esperan que la cumbre de Río adopte compromisos claros en ese sentido. Fue él quien lanzó en 1997 la idea de reunir en una cumbre a los líderes de América Latina y Europa. La diplomacia española ha mantenido un breve y discreto silencio sobre el enfrentamiento que se registró el pasado viernes en Colonia, Alemania, entre el presidente del Gobierno español, José María Aznar, y el presidente francés, Jacques Chirac, sobre el mandato de negociación con Mercosur. Aunque la cumbre es mucho más que Mercosur -asistirán 17 Jefes de Estado y de Gobierno de Latinoamérica, 16 del Caribe y 15 de la UE-, las negociaciones con el bloque en el que están Brasil y Argentina son el escaparate de las relaciones de Europa con el resto del continente." />
    <Subject IPTC= " Politit & Treaty and International Organisations & alliances" />
    <Document>Francia trata de retrasar la negociación de un pacto comercial de la UE con América Latina.
    </Document>
    <Document>El portazo de la UE a Mercosur amenaza el futuro de las relaciones con América Latina.</Document>
    <Document>América Latina busca en la UE la salida a las crisis financieras cíclicas.</Document>
  </Event>
  - <Event Id= "Cluster_343" Level= "1">
    <Subject IPTC= "Politit & Treaty and International Organisations & alliances" />
    <Document>Francia culpa a España del bloqueo de la negociación entre la UE y Mercosur.</Document>
  </Event>
</Topic>

```

Fig. 2. A snippet of the created hierarchy.

Moreover, our system has been evaluated using the TDT2 dataset, version 4.0. This corpus consists of 6 months of news stories from the period January to June

Table 2. Evaluation of the obtained subjects.

Topic	Subject	Eval.	Topic	Subject	Eval.
Elections in South Africa	Government /President	R	The murder of the bishop Gerardi	Crime/Homicide	G
Ocalan's trial	Justice/Trials	G	Electric belt in USA	Economy & Finance /Clothing	B
Cachemira's conflict	Conflicts and war	G	Earthquake in Mexico	Disaster/ Earthquake	G

1998. The news stories were collected from six different sources. Human annotators identified a total of 193 topics in the TDT2 dataset. 9824 English stories belong to one of these topics, the rest are unlabeled. The collection TDT2 is not annotated in different abstraction levels. Therefore we only have the manual classification at Topic Level. Again, we build a hierarchy of two levels.

Table 3 shows the minimum topic-weighted and normalized costs and the maximum F1-Measure for our system on the TDT2 dataset. Again, the better results are obtained when we use our similarity measure. We consider that the results obtained by *JERARTOP* are very good if we compare them with those obtained by the existent detection systems (see evaluation results in [11]).

Table 3. The obtained results in TDT2 collection.

System	Cost	Norm Cost	F1-Measure
Base cosine	0.0047	0.235	0.730
<i>JERARTOP</i>	0.0054	0.2688	0.779

6 Conclusions

In this paper we presented a new detection system. The most important novelty is the construction of a hierarchy of topics and events with different abstraction levels. This hierarchy allows users to discover the temporal structure of topics and events, that is, to identify not only the topics but also the possible smaller events they comprise. Unlike the detection systems proposed in the literature, this approach avoids the definition of what a topic is, since is the user who navigating by the created hierarchy decides in what level he explores the knowledge of interest.

A new similarity measure between documents considering the places, the temporality and contents of the news is also introduced.

Additionally, we propose two description methods of the obtained clusters at any level of the hierarchy. Thus, *JERARTOP* categorizes each node of the created hierarchy on the pre-specified subject matters and it provides a summary of the associated contents in each one of them. These methods are helpful to a user in order to determine at a glance whether the content of an event is of interest.

The obtained results for the F1-measure and the detection cost also demonstrate the validity of our algorithms for topic detection tasks. *JERARTOP* not only obtains similar results to the best systems in TDT2 evaluation but rather it also offers additional information: the hierarchy of topics and events along its description.

References

1. Allan, J.; Carbonell, J.; Doddington, G.; Yamron, J. and Yang, Y.: Topic Detection and Tracking Pilot Study: Final Report. In *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, pp. 194-218, 1998.
2. National Institute of Standards and Technology. The Topic Detection and Tracking Phase 2 (TDT2) evaluation plan. version 3.7, 1998.
3. Carbonell, J.; Yang, Y.; Lafferty, J.; Brown, R.D.; Pierce, T. and Liu, X.: CMU Report on TDT-2: Segmentation, detection and tracking. In *Proceedings of DARPA Broadcast News Workshop*, pp.117-120, 1999.
4. Yamron, J.: Dragon's Tracking and Detection Systems for TDT2000 Evaluation. In *Proceedings of Topic Detection & Tracking Workshop*, pp.75-80, 2000.
5. Allan, J.; Lavrenko, V.; Frey, D. and Khandelwal, V.: UMASS at TDT 2000. In *Proceedings TDT 2000 Workshop*, 2000.
6. Brants, T.; Chen, F. and Farahat, A.: A System for New Event Detection. Annual ACM Conference on Research and Development in Information Retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, SIGIR'03, Toronto, Canada, 2003.
7. Pons-Porrata, A.; Berlanga-Llavori, R. and Ruiz-Shulcloper, J.: On-line event and topic detection by using the compact sets clustering. *Journal of Intelligent and Fuzzy Systems*, Number 3-4, pp.185-194, 2002.
8. Pons-Porrata, A.; Berlanga-Llavori, R. and Ruiz-Shulcloper, J.: Building a hierarchy of events and topics for newspaper digital libraries. *LNCS 2633*, Springer-Verlag, 2003.
9. Pons-Porrata, A.; Ruiz-Shulcloper, J. and Berlanga-Llavori, R.: A method for the automatic summarization of topic-based clusters of documents. *LNCS 2905*, Springer Verlag, pp. 596-603, 2003.
10. Pons-Porrata, A.; Berlanga-Llavori, R. and Ruiz-Shulcloper, J.: A new word sense disambiguation method using WordNet. *Proceedings of X Conference of the Spanish Association for the Artificial Intelligence*, Volumen II (In Spanish), pp. 63-66, 2003.
11. Fiscus, J.G.; Doddington, G.; Garofolo, J.S. and Martin, A.: Nist's 1998 topic detection and tracking evaluation (tdt2). In *Proc. of the DARPA Broadcast News Workshop*, 1998.