A News Domain Topic Detection System

CORMAC FLYNN and JOHN DUNNION
Intelligent Information Retrieval Group,
Department of Computer Science
University College Dublin
Dublin
IRELAND

In this paper we describe a system that performs Topic Detection, a sub-task of the Topic Detection and Tracking (TDT) Project. We describe the Topic Detection task and present initial results for both a baseline system and a set of extensions that we have implemented in our system that attempt to model events and reportage in the news domain. We describe how our system clusters documents from a TDT corpus and from live news feeds and presents this data to the user in a variety of formats. We conclude that our system produces interesting and useful clusters, and outline some areas of future work.

Keywords: Online news, topic detection, document clustering.

1 Introduction

From 24-hour news networks and the ever-expanding print and broadcast media, to online news outlets that deliver live reportage daily from around the world, news reportage is now all-pervasive and instant, available in multiple languages and to an international audience. The proliferation of sources has introduced a number of profound problems. For the average user, it is increasingly difficult to locate and retrieve complete and comprehensive information about an incident. Time constraints mean users often limit themselves to a handful of sources, offering only a narrow view on events. To follow a news event in its totality, from initial report through each twist and development to its conclusion, is an increasingly difficult task. Moreover, traditional "search and retrieve" techniques are ill-suited for general queries such as "What happened?". Because of these problems, users are increasingly seeking a single channel for their news and information needs.

The *Topic Detection and Tracking (TDT) Project* is an attempt to provide such a source. The aim of the TDT Project is to provide language- and platform-independent technologies to monitor sources of news reportage, detect breaking stories and track these as they develop over time. Research in this area began with a pilot study in 1997 [1] that defined the problem, outlined the tasks that a TDT system would be required to perform and established an initial corpus of news articles, extracted from Reuters newswire and CNN broadcasts. Since then, there have been annual open evaluations during which the required tasks have been refined and the input corpora expanded and enriched with content from a variety of multi-lingual sources. Currently, there are five TDT tasks: *Story Segmentation*, *New Event Detection*, *Topic Detection*, *Topic Tracking* and *Link Detection*. In our research we have concentrated on the Topic Detection task, ie the grouping together of stories that discuss the same event into topical clusters. This can be performed on either a retrospective corpus of documents or on a live stream. A complete Topic Detection system would allow the user to gather news from a range of sources, producing a set of clusters that represent the events that have occurred in the time period covered. This task differs from standard document clustering, where the objective is to group topically related documents into clusters that capture general categories or topics. For our purposes, we define a topic over a corpus to be a set of documents that share a consistent theme or concept. Two documents can lie in the same topic yet still cover different specific issues, eg a news article on a forest fire and one that reports on an earthquake are both members of the topic "Natural Disasters". It is possible to imagine any number of equally valid topic boundaries for a particular dataset. For Topic Detection, we aim for clusters that reflect the full narrative of an event as it grows and develops over time. Unlike a set of topics, there are a finite number of valid events that could take place for a collection of TDT documents. Furthermore, we are clustering documents taken from a single specific genre, that of news reportage.

In this paper, we describe the design of a baseline system for Topic Detection. We outline a set of domain-informed extensions that attempt to produce clusters that better represent an event narrative, and present some initial results. Finally, we show how our system extracts and clusters live reportage from online RSS feeds and demonstrate its ability to produce coherent and comprehensive event clusters from a variety of sources.

2 Overview of TDT Project

2.1 TDT Tasks

As stated above, the TDT pilot study was carried out in 1997. This study described three tasks that an eventual

Topic Detection and Tracking system would be required to perform: *Story Segmentation*, *Topic Detection* and *Topic Tracking*. In the most recent phase of the TDT evaluation, these tasks were further expanded and refined to include *New Event Detection* and *Link Detection*.

Story Segmentation is the division of an audio or visual stream into distinct stories, using either the original data or a textual transcript. The goal is to define boundaries between contiguous stories in a continuous data stream.

Topic Tracking is the detection of news articles that discuss particular target topics. Incoming stories in a data stream are associated with events already known to the system. An integrated TDT system would allow the user to highlight those stories in which they have an interest, and then to track all on-topic reports surrounding this event as it develops over time.

Topic Detection refers to the detection of events in terms of the stories that discuss them. In contrast to the tracking task, the system has no previous knowledge of the detected events. The detection task can operate on a retrospective corpus of news articles or on an online stream. In the retrospective case, an entire set of documents is grouped into topical clusters that match the events that occurred during the period covered by the corpus. In the online case, the user would be alerted to news events as they arrive in realtie on the stream.

New Event Detection is the identification of the initial story that discusses a topic, alerting the user when a new or previously unseen event has occurred.

Link Detection is the process of deciding whether a pair of stories discuss the same topic or are linked by a common event. Link Detection is a component technology required for the other tasks.

In our work, we have concentrated primarily on the Topic Detection task. Our initial research indicated that standard Information Retrieval techniques combined with domain-informed extensions might usefully be applied to this task. Furthermore, the consistent and relatively error-free nature of the TDT-1 corpus allowed us to design, implement and evaluate our system quickly and to test some of our assumptions about the news domain.

2.2 TDT-1 Corpus

The TDT-1 pilot study established both an initial corpus and a set of evaluation tools, designed to benchmark the performance of competing TDT systems in a standard way. The TDT-1 corpus comprises 15863 documents, both standard news reports and manually transcribed broadcast news, taken from Reuters newswire and the CNN news network. This data was gathered over a 12-month period, from 1 July 1994 to 30 June 1995, and marked-up in Standard Generalised Mark-up Language (SGML). SGML is a standard mark-up language for the description of text, used to define the structure and contents of a document and to ensure that the data is portable and comprehensible to humans. Each document in the TDT-1 corpus was marked up with SGML tags, describing the information in each part of the document, eg title, date, dateline, news source, speaker, etc.

The TDT-1 corpus is arranged chronologically, beginning on 1 July 1994. Unlike more recent TDT corpora, the TDT-1 corpus contains only English-language sources, does not include automatically transcribed broadcast information and contains only textual data.

In order to evaluate the performance of a system operating on the TDT-1 corpus, 25 events were chosen, amounting to approximately 7% of the documents. Each document in the corpus was then manually annotated in terms of these target events. A news article that was directly related to a particular story was labelled with a **YES** judgement. A report that contained no mention of the target event was labelled with a **NO** judgement. For those documents which mentioned the story in passing (approximately 10% of the total corpus), the **BRIEF** label was used. The **BRIEF** label was ignored in many of the task evaluations and was dropped from later revisions of the TDT corpus.

On average, each of the labelled events contained 55 documents and spanned around 121 days. The longest event was the “DNA in OJ Trial” event, which spanned 353 days, and the shortest was “Cessna on White House”, which took place over a single day. The “OK City Bombing” event contained the greatest number of directly on-topic documents, with 295 news reports labelled with the **YES** tag, while the event with the smallest number of on-topic documents was the “Karrigan/Harding” event, with 3 relevant articles. For all but two of the target events, the initial report for each event occurred in the period covered by the corpus.

The Topic Detection task uses the full TDT-1 corpus. For retrospective Topic Detection, the complete corpus is taken as input and a partitioning of the dataset is produced as output, where each story cluster corresponds to a distinct event. For online detection, the corpus is analysed as a stream, to simulate the real-time requirements of an online Topic Detection system. A list of YES/NO confidence scores is also generated, indicating whether the system deems an article to be the beginning of a new event at the time of its arrival.

2.3 TDT Evaluation

The TDT-1 evaluation includes a set of tools that provide a standard way to compare systems operating on the TDT-1 corpus. The metrics used to measure performance include *precision* and *recall*, as traditionally used in information retrieval systems. Additionally, Topic Detection and Tracking systems are also evaluated in terms of miss rate and false alarm rate. A *miss* occurs when the system overlooks a document that is relevant to a particular event. Conversely, a *false alarm* occurs when an article is erroneously flagged as being relevant to a target story.

There are advantages and disadvantages to both sets of measurements. Recall and precision tend to be more

	Relevant	Not Relevant
Retrieved	A	B
Not Retrieved	C	D

$$Recall(R) = \frac{A}{A+C}$$

$$Precision(P) = \frac{A}{A+B}$$

$$Miss = \frac{C}{A+C}$$

$$FalseAlarm = \frac{B}{B+D}$$

$$F1 = \frac{2PR}{P+R}$$

Figure 1. *Formulæ for the TDT-1 performance measures.*

useful for measuring the performance of applications, since high recall and precision rates produce clear and measurable results for the user. However, for statistically evaluating a system, precision in particular is very often not sufficiently fine-grained to adequately capture the actual performance of a system or the impact of individual modules on the system's overall effectiveness. For this reason, the additional measures of miss and false alarm rates are chosen to better highlight any potential performance improvements in a developing system. Furthermore, this method of evaluation can be used to graph the performance of a TDT system on a *Detection Error Trade-off (DET) curve*. A DET Curve plots the performance of a system in terms of its miss and false alarm rates. Since we wish to minimise both these performance measures, curves lying close to the origin indicate a system that performs well. DET curves can be generated automatically by the TDT-1 evaluation tools and offer a standard way to graphically represent the performance of different TDT systems.

We can also produce a single-value score, known as the *F1 measure*. The F1 measure ranges from 0 to 1, and can be used to compare the general performance of competing TDT systems. The formulæ for recall, precision, miss rate, false alarm rate and F1 measure are given in Figure 1.

The TDT-1 evaluation software uses these measures to judge the average performance of a TDT system. Recall, precision, miss rate, false alarm rate and F1 measure may all be calculated under two headings. The *micro-average*, or *pooled-average*, method merges the overall performance scores for each of the 25 target events and uses these results to produce global performance measures. The *macro-average* approach produces a single performance value for each of the 25 labelled events first, and then takes the average to give an overall performance result. Because of the small number of labelled events in the TDT-1 corpus, the micro-average results are typically favoured, although both are usually given.

3 Baseline Topic Detection System

The baseline system is composed of three modules: the Pre-processing module, the Clustering module and the Presentation component. The baseline system accepts documents from either the TDT-1 corpus or live RSS feeds.

The Pre-processing module builds a representation for each document using the smallest set of terms possible, as defined by our feature selection criteria. A feature is chosen according to the following criteria:

- global frequency;
- local frequency;
- document frequency;
- whether the word is a stop-word.

Terms that appear infrequently or in only a few documents are sometimes excluded at this stage to reduce the dimensionality of the term vectors. However, for the news domain, such words are often representative of short events or incidents for which there are only a small number of relevant documents. Moreover, the phenomenon of topic shift [2], where the focus of an event changes suddenly, perhaps signalling a new or unexpected development, means that previously infrequent terms can increase in both frequency and importance. For this reason, we choose terms irrespective of their local or document frequency. For global frequency, we exclude only those words that occur once across the entire corpus, since these are unlikely to be representative of a document. Once these terms have been collected, the final pre-processing step is to apply appropriate weights.

The Clustering module takes this set of weighted features and groups the document collection into event clusters. Document clustering techniques fall into one of two categories. Partitional Clustering starts with all items in a single cluster and attempts to split this until k disjoint clusters have been formed. Hierarchical Clustering begins with each item in its own singleton cluster and aims to merge the most similar pair until the desired number of k clusters is reached. Partitional methods tend to be fast, whereas hierarchical approaches tend to produce better clusters. We attempt to combine the advantages of both methods by implementing a clustering approach similar to that outlined in [3], where a slow accurate hierarchical method is combined with a fast partitional algorithm. We examined three agglomerative (bottom-up) hierarchical algorithms: single-link, complete-link and group-average. Group-average techniques have proven useful in previous TDT research [4, 5], and this was borne out by our evaluation.

The Presentation module displays each cluster to the user. This requires some representation of the cluster's

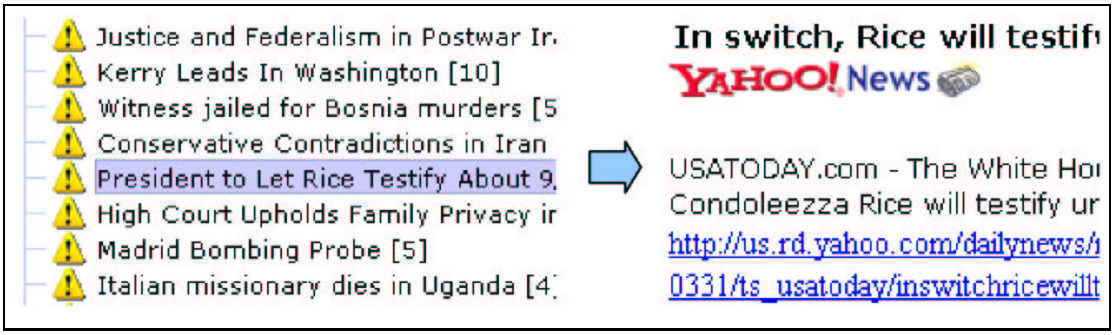

Figure 2. *A partial cluster tree and document for an RSS feed.*

overall contents. We achieved this by first finding the centroid, ie the normalised average term vector for all members of the cluster excluding outliers, and by representing the cluster contents by the document that lies closest to this centroid. In the user interface, this is shown as a tree, where each node is the title of the representative document for a particular cluster. Figure 2 shows part of a cluster tree for an RSS feed and a document from the feed.

4 Domain-informed Topic Detection

The TDT project is concerned with documents drawn from the specific domain of news reportage. Moreover, we wish to group these documents into clusters that represent the events that have occurred for the period covered. We attempted to enhance our baseline system with a number of extensions that better exploit the particular characteristics of the news domain, both the common discourse of news reportage and the typical distribution of documents for an event in a temporal data stream [6, 7].

Events have a beginning and an end, a date we associate with the occurrence and a time by which the incident has played out in full. Between these two points, there is a broad pattern of development common across news events. The initial stages of an event are characterised by a flurry of directly on-topic news reports. Since the reader is still unfamiliar with the details, these reports typically contain more directly relevant information than later articles. As time passes and the event continues to develop, this burst of articles starts to drop off. This is represented both by a decrease in the number of articles and an increase in the time gap between successive stories. This pattern can be seen consistently in the histograms for some of the events in the TDT-1 corpus. Moreover, if we examine Google search terms for a particular news event, a similar distribution can be found (see Figure 3).

Moreover, common patterns can also be observed for the news documents themselves. News articles share an "inverted triangle" layout, with a headline and a lead paragraph followed by a main body of text, where the value of the information decreases as we move through the document. The important details of an event, the *who*, *when* and *where*, tend to occur in the headline and opening paragraph [8, 9]. The main body of the report expands on the information set out in the lead, providing background details, quotes, analysis or offering different perspectives on the incident, but rarely containing items more newsworthy than those in the lead.

For our first set of system extensions, we employ a dynamic threshold that is higher for the early stages of an event cluster. This approach is based on the assumption that, since earlier documents contain more directly on-topic information, they are likely to be more similar to one another than later articles. Using a higher threshold discourages spurious merges early in the clustering process. Secondly, an alternative *incremental* approach to clustering is used, that restricts inter-cluster comparisons to those that fall within an adaptive time window. This window looks further and further ahead in time as the period covered by the cluster increases. This is an attempt to model event distribution, where on-topic articles decrease in number and increase in distance as the incident develops.

For the common discourse of news articles, we added an extension that restricted indexing to a predefined percentage of the text, starting from the top down, ie the top 10%, the top 20%, etc. Secondly, we implemented an alternate term weighting scheme that favours words in the upper half of the document.

All of the above system extensions aim to focus indexing and clustering on the information that is likely to have the highest value. Although experiments are ongoing, Table 1 shows some initial results for both the baseline and the discourse extensions, compared against similar TDT-1 retrospective Topic Detection systems. Using this extension alone, we observed a 13% increase in recall over the baseline with only a 7% drop in precision. Moreover, the system outperforms the best result for the TDT-1 evaluation.

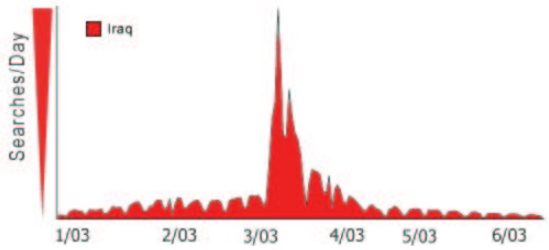

Figure 3. *Histogram for Google searches surrounding the Iraq (2003) conflict.*

Run	%Miss	%FA	%Recall	%Prec	Micro F1	Macro F1
CMU	38	.04	62	82	.71	.79
UMass	66	.09	34	53	.42	.60
Dragon	39	.08	61	69	.65	.75
Baseline	38	.04	62	80	.70	.80
Baseline + Exts	**25**	**.08**	**75**	**73**	**.74**	**.80**

Table 1. *Topic Detection results compared with the TDT-1 Pilot Study.*

5 RSS Feed Clustering

Rich Site Summary (RSS) is an XML-based meta-language that is increasingly used to distribute online syndicated data. We developed a version of our Topic Detection system that gathers news articles from RSS sources and groups them into clusters representing the most recently occurring events. RSS is lightweight enough and our clustering algorithm sufficiently fast that a large number of news items can be clustered in only a few seconds. This produces a set of event clusters that gathers reportage from sources that differ greatly in detail, style and content, offering a variety of perspectives on the same incident. Moreover, this data can be viewed both as a list of documents and as a histogram showing the event's distribution over time. The system architecture is such that this set of visualisations could easily be expanded as part of our future work, to include a more detailed timeline of occurrences or even to place events on a map interface. Although we have still to perform a full analysis of our online system, we found that even this basic implementation produced useful and interesting clusters and offered a coherent and comprehensive view of events.

6 Conclusions and Future Work

We have described the design of a baseline Topic Detection system and outlined a set of domain-informed extensions that better model events and reportage in the news domain. Although experiments are ongoing, initial results are promising. Furthermore, we have described how our system can be applied to news from live RSS feeds to produce useful clusters that offer a coherent and comprehensive view of events. As part of our future work, we intend to further develop both the clustering algorithm itself and the online implementation. We also intend applying our system to the new TDT-4 corpus and participating in the official TDT evaluation.

Acknowledgements

The support of the Enterprise Ireland Informatics Research Initiative is gratefully acknowledged. The research was funded under grant PRP/00/INF/06.

References

[1] James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. Topic Detection and Tracking Pilot Study. In *Proceedings of the DARPA Broadcast News Workshop*, 1998.

[2] Avi T. Arampatzis, Th.P. van der Weide, C.H.A. Koster, and P. van Bommel. Term Selection for Filtering based on Distribution of Terms over Time. In *Proceedings of the 6th Conference on Content-Based Multimedia Information Access (RIAO 2000)*, pages 1221–1237, Paris, France, April 2000.

[3] Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey. Scatter/Gather: a Cluster-based Approach to Browsing Large Document Collections. In *Proceedings of the 15th International ACM SIGIR Conference (SIGIR 1992)*, pages 318–329, 1992.

[4] Vasileios Hatzivassiloglou, Luis Gravano, and Ankineedu Maganti. An investigation of linguistic features

and clustering algorithms for topical document clustering. In *Proceedings of the 23rd International ACM SIGIR Conference (SIGIR 2000)*, pages 224–231. ACM Press, 2000.

[5] Yiming Yang, Jaime Carbonell, Ralf Brown, Tom Pierce, Brian T. Archibald, and Xin Liu. Learning Approaches for Detecting and Tracking News Events. *IEEE Intelligent Systems*, 14(4):32–43, 1999.

[6] Cormac Flynn and John Dunnion. Domain-informed topic detection. In *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2004)*, Lecture Notes in Computer Science, Volume 2945, pages 617–626. Springer-Verlag, 2004.

[7] Cormac Flynn. Topic detection in the news domain. MSc thesis, University College Dublin, 2004.

[8] T.A. van Dijk. *News as Discourse*. Lawrence Erlbaum, Hillsdale, NJ, 1988.

[9] Allan Bell. *The Language of News Media*. Blackwell Publishing, Oxford, 1991.