Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster

Kathleen R. McKeown Regina Barzilay David Evans Vasileios Hatzivassiloglou
Judith L. Klavans Ani Nenkova Carl Sable Barry Schiffman Sergey Sigelman
Department of Computer Science
450 Computer Science Building
Columbia University
1214 Amsterdam Avenue, New York, N.Y. 10027
{kathy, regina, devans, vh, klavans, ani, sable, bschiff, ss1792} @cs.columbia.edu

ABSTRACT

Recently, there have been significant advances in several areas of language technology, including clustering, text categorization, and summarization. However, efforts to combine technology from these areas in a practical system for information access have been limited. In this paper, we present Columbia's Newsblaster system for online news summarization. Many of the tools developed at Columbia over the years are combined together to produce a system that crawls the web for news articles, clusters them on specific topics and produces multidocument summaries for each cluster.

General Terms

Multidocument summarization

Keywords

clustering, topic detection and tracking, Columbia Newsblaster

1. INTRODUCTION

Current technology in summarization and topic detection and tracking is mature enough to be used reliably in a live, online environment. Newsblaster is a system developed at Columbia to provide news updates on a daily basis; it crawls news sites, filters out news from non-news (e.g., ads), groups news into stories on the same event, and generates a summary of each event. Summaries are generated using the Columbia Summarizer [8, 7, 13], which was evaluated in the Document Understanding Conference (DUC) in 2001. News is grouped into stories on the same event using a Topic Detection and Tracking (TDT) style system developed at Columbia [4]. Unlike other TDT systems, Columbia's uses a learned, weighted combination of features to determine similarity of stories, grouping articles on the same event. Newsblaster (`http://www.cs.columbia.edu/nlp/newsblaster`) typically generates summaries on clusters of between five and 30 news stories, but in cases where an event generates a lot of interest, we may find a much larger cluster; for example, on November 23rd, Newsblaster generated a summary of 75 articles all describing the advance of troops towards Kunduz. We began running Newsblaster on September 16th, 2001, when the summarization system was robust enough to run on a daily basis. We were also interested in archiving news on the September 11th events. The archives (see the bottom of the Newsblaster web page) show the progress on the interface since that time, as we have moved from very rudimentary classification of events to a three-level hierarchical classification in the current version.

Newsblaster is unique in its integration of TDT and summarization to provide a system for daily browsing of news. Integration of the separate components, for which papers have previously been published, introduced a number of unexpected research issues. First, one of our summarizers had been designed to generate summaries on a single event. Current TDT approaches produce clusters that are topically related, but not necessarily on a single event. We needed to create fine-grained clustering that could produce the kind of input needed by the summarizer. Second, given the wide variety in news that we find in an online environment, we needed multiple strategies to allow us to summarize different kinds of news clusters. This led to our use of multiple summarization systems within one architecture, with a router which automatically determines which summarization system should be invoked. Third, the user interface needs inherent in displaying large quantities of information each day created demands on the clustering and summarization process. For example, in our early implementation, we used a single level of clustering to feed document sets to the summarizer. However, the results did not provide a very useful breakdown of large quantities of news for browsing. This led to our interleaving of categorization with clustering to create a hierarchical view of the news along with a new approach to generating labels of clusters.

In this paper, we overview the system, discussing the gathering of news, the organization into events, subsequent summarization of events, and the integration of related images.

2. SYSTEM DESCRIPTION

Newsblaster follows a pipeline architecture. First, the system crawls the web for news articles, followed by a pre-processing phase which normalizes the text into a standard form and extracts images. The news articles are clustered into event clusters, then these clusters are grouped at a higher level, putting together related

HLT '02, San Diego, Calif., USA
.

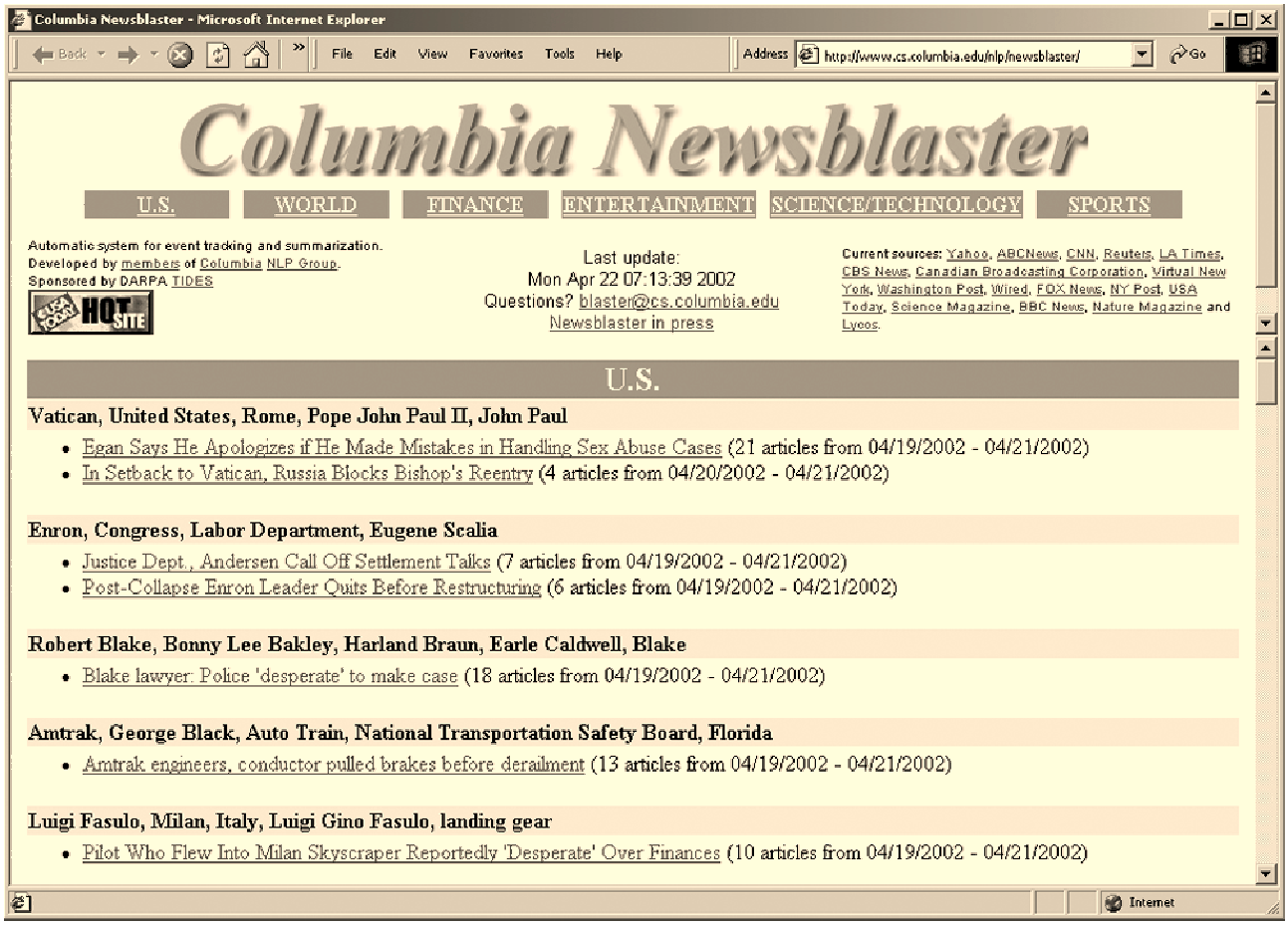

Figure 1: Newsblaster frontpage

events. Events are also categorized into one of the six top-level categories. A multidocument summary is created for each event cluster and augmented with pictures extracted from the articles.

3. GATHERING NEWS

Newsblaster currently crawls 17 news sites including those of CNN, Reuters, Fox News, NY Post, and USA Today, among others. The list of sites is stored in a text file and may change over time. Each site is traversed up to a maximum depth (currently 4) and only links within the site are considered. For each page examined, if the amount of text in the largest cell of the page (after stripping tags and links) is greater than some particular constant (currently 512 characters), it is assumed to be a news article, and this text is extracted. Note that tables are used in many web sites to format the page and the largest table cell usually contains the body text of the article.

4. ORGANIZATION OF STORIES

Newsblaster hierarchically classifies the news stories gathered by the crawler into three levels. At the top level, it uses text categorization to determine whether the story falls into one of six predetermined categories: US, International, Financial, Entertainment, Science and Technology, or Sports. We calculate category TF*IDF vectors for each category, and compare the TF*IDF vector for each article in a given cluster to all categories. In this manner, articles are assigned to the closest (most similar with a cosine measure) category. Then we classify the cluster into the category to which the largest number of articles in the cluster are assigned. During these calculations, we smooth estimated frequencies in the articles using smoothing bins [10]. These categories are shown in large orange bands on the main Newsblaster front page, shown for the US category on April 22nd, 2002 in Figure 1.

Within each broad category of articles, Newsblaster further organizes the news stories into two hierarchical levels. The lowest level corresponds to articles on the same event (see the underlined text as in"*Egan Says He Apologizes if He Made Mistakes in Handling Sex Abuse Cases*", Figure 1), while the higher level groups together related events (shown in bold, heading each group of events as in "*Vatican, United States, Rome, Pope John Paul II*", Figure 1). For both of these, we use a hierarchical clustering system developed at Columbia [4]. On this particular day, there were two specific events related to the Vatican and the United States. While related, each of the clusters is a different event which has its own summary page containing the articles about the event.

The system uses agglomerative clustering with a groupwise average similarity function. It is distinguished by its use of not only the usual TF*IDF weighted words, but also linguistically motivated features, such as terms, noun phrase heads and proper nouns, likely to correlate with events and not simply topically related stories.

It also incorporates a log-linear statistical model for automatically adjusting the relative weights of the different features. We

Figure 2: A MultiGen Summary

have evaluated our news clustering system using TDT-2 data. Our system performs comparably to the top TDT-2 participants under most measures, and outperforms them in macro-averaged detection cost [4]. For its use within Newsblaster, we have empirically determined two thresholds for the clustering at each of the two levels (single event and group of related events).

To facilitate the user's interaction with the categorized stories, we also provide labels for each cluster. For the lowest (event) level, where the articles are closely related in content, we use heuristics to select the article that is most related to the other articles in each cluster, and label the entire cluster with that article's title (shown in blue on the main page). For the second level (related events) we extract from all articles in the cluster all proper names and terms, then weigh those according to their total frequency and inverse document frequency, and select up to five such terms that are most representative of the related events (shown in black, heading each group of events).

5. SUMMARIZING EVENTS

All sets of clustered articles corresponding to the lowest event level are sent to the Columbia Summarizer to generate summaries of events. The Columbia Summarizer is a composite summarization system that uses different summarization strategies dependent on the type of documents in each cluster; this contrasts with other systems which typically perform one type of summarization only [3]. A router automatically determines the type of documents in each cluster and invokes the appropriate summarization subcomponent.

Using the training corpus provided for DUC, we manually derived a typology of document sets. *Single-event* documents center around one single event happening at one place and at roughly the same time, involving the same agents and actions. *Person-centered* (or "biography") documents deal with one event concerning one person and include background information about that person. *Multi-event* documents describe several events occurring at different places and times, and usually with different protagonists, are reported together. There is a common theme to these events, e.g., a cluster might collect many fire incidents on unrelated cruise ships. The time span covered is unpredictable, but longer than in the single-event case. *Other* clusters contain even more loosely related documents and do not fit any of the categories above.

To summarize documents on the same event, the Columbia summarizer uses an enhanced version of MultiGen [2, 8]. MultiGen integrates machine learning and statistical techniques to identify similar sentences (set set of similar sentences is called a *themes*) across the input articles [5, 6]. It then uses an alignment of parse trees to find the intersection of similar phrases within sentences [2]. It orders the selected themes [1] and uses language generation to cut and paste together similar phrases from the theme sentences. Each theme corresponds to roughly one sentence of the summary.

For biographical documents, it uses an alternate system, DEMS (Dissimilarity Engine for Multidocument Summarization) [13], tuned to the biographical task; and for sets of loosely similar documents, it uses DEMS with a more general configuration. DEMS selects

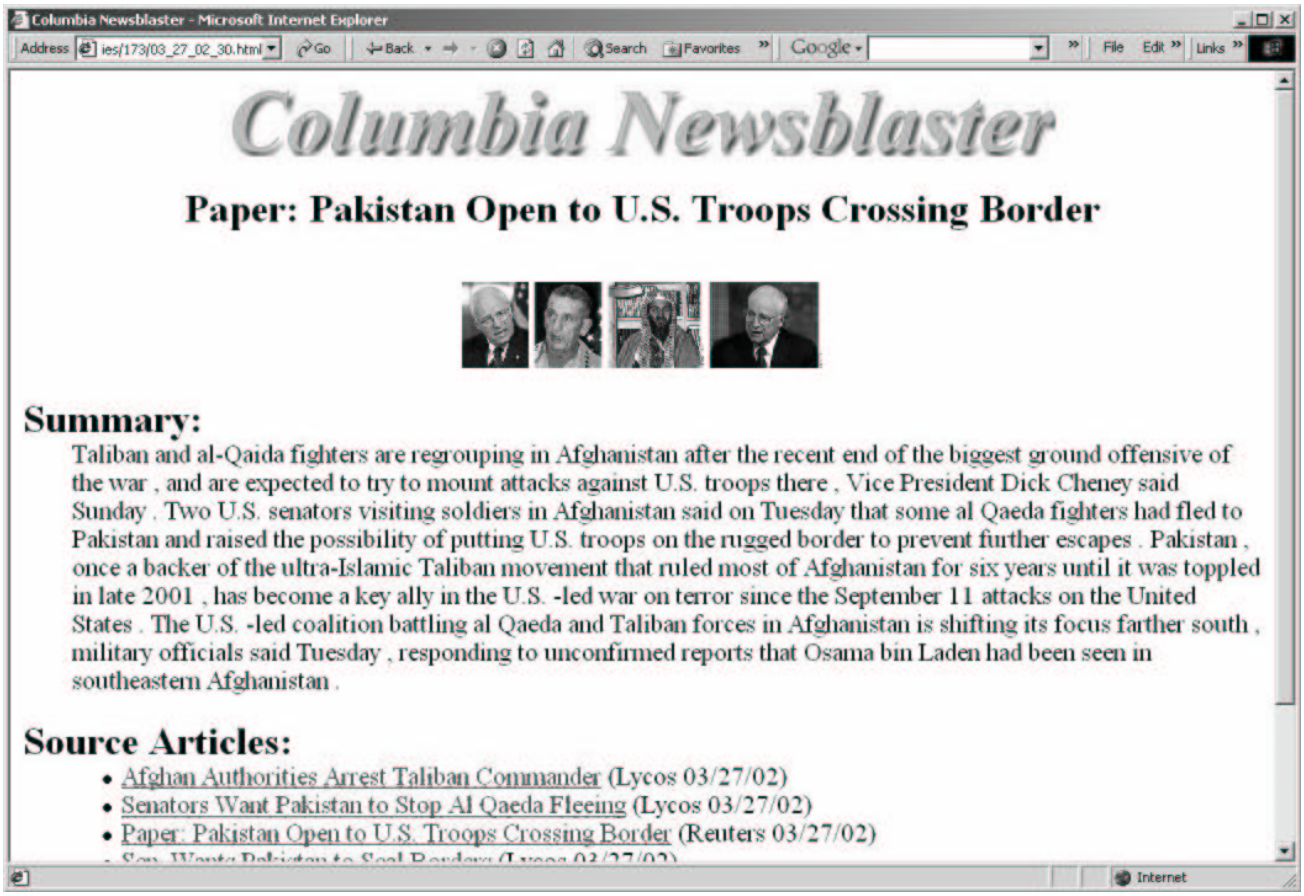

Figure 3: A DEMS summary

sentences that contain information that is important or interesting enough to be included in a summary. It uses a combination of several features that are critical for new-information detection with some traditional heuristics used in single-document summarization [13]. For example, it uses frequencies of concepts (or sets of synonyms) as opposed to individual words, combined with global information about what words are likely to appear in a lead sentence, to determine whether an article sentence should be included in the summary. For biographies, it also incorporates techniques used in the BioGen system, developed jointly by Mitre and Columbia [12]. Since there were no sets of similar sentences in the input, the generation strategy that we used in MultiGen was not applicable in DUC. We have investigated some ways to rewrite the summary [13] and are continuing to explore techniques.

While all clusters in Newsblaster are more closely related than in DUC, we still needed a summarizer to handle clusters that are not as closely related as most single event descriptions. For example, Figure 3 shows a summary of a loosely connected set of articles related to Afghanistan, Pakistani borders with Afghanistan and a recent visit by Cheney to the area. DEMS has selected different important facts from the articles, which are not necessarily sequentially or strongly linked to each other (Figure 3. In this case, the summary provides an overview of the different topics touched on in the articles.

Figure 2 shows a MultiGen summary on a single event, a fire that grew out of control in New Mexico. In this summary, all sentences relate different stages in the sequence of events, with the exception of the sentence on fires last year. These sentences are ordered chronologically in the summary. Individual summary sentences have been drawn from pieces of several article sentences. Figure 6 shows how the first sentence of the summary was generated from pieces of similar phrases in the theme from which it was produced. Here, the third theme sentence provides the summary wording on how the blaze was started. Note that while the exact wording from the third sentence is used, this information is repeated across the different sentences of the theme. The second sentence of the theme contributes the clause "*mistakenly thinking the ashes were cold.*". This information is repeated across two sentences of the theme.

6. ADDING RELATED IMAGES

Newsblaster selects and displays thumbnails of images that are related to an event on the same page where the summary of events is displayed. During the web crawling phase, in addition to looking for news articles, Newsblaster also looks for embedded images in the articles. Since articles are taken from many different sources on the web, and these sources might change over time, it is important that the rules used by the system to find such images are general, and can be applied to multiple sites. The rules must extract most of the appropriate images without also taking advertisements and other inappropriate images. We weighted precision higher than recall, since users will not notice if certain images are not found, but inappropriate images would be visible. Some patterns we noticed

when manually examining news sites were: (1) Images are almost always in the same cell as the article or an embedded cell. (2) Images that are jpeg's tend to be appropriate, but other formats are more likely advertisements or link related. (3) Images with a word like "ad" or "advertisement" in the URL are probably not appropriate. By combining such rules, our system seems to achieve nearly perfect precision while still recalling the high majority of appropriate images.

We are currently working on rules to retrieve corresponding captions to the images that have them. In prior research, we have developed techniques for categorizing images based on their captions [11] as well as image features [9]. In current, ongoing research, we are exploring the categorization of images into sub-categories of news categories based on shallow parsing of captions and simple word similarity metrics. For example, for news concerning disasters, we have found that the main subject and verb from the first sentence of an image caption is often enough to determine if the image likely focuses on victims, workers responding, or wreckage. This research can be incorporated in Newsblaster by displaying images with similar content together, allowing users to further browse or search the group.

Figure 4: Theme for the first sentence of the MultiGen summary

<Gov. Gary Johnson said> <the New Mexico blaze> <started when a resident dumped fireplace ash in a back yard> <mistakenly thinking the ashes were cold.>

<**THEME 11**>

<Art 1 Par 1 (global para num: P1) >
Fireplace ashes dumped in a back yard sparked a grass and timber wildfire that burned 28 homes in an affluent neighborhood in the mountains of southern New Me xico, authorities said Sunday.

<Art 2 Par 14 (global para num: P50) >
Gov. Gary Johnson, who toured the fire zone Sunday, said the blaze started Saturday when a resident dumped fireplace ash in the back yard, <**mistakenly thinking the ashes were cold.**>

<Art 3 Par 7 (global para num: P72) >
<**Gov. Gary Johnson said**> the south - central <**New Mexico blaze**>, which charred about 960 acres, <**started when a resident dumped fireplace ash in a back yard**> in the mistaken belief the ashes were cold.

<Art 7 Par 11 (global para num: P140) >
Gov. Gary Johnson said Sunday the fire was sparked when a local homeowner dumped what he believed were harmless fireplace ashes into his backyard.

7. CONCLUSIONS AND CURRENT DIRECTIONS

This paper summarizes innovative contributions in the areas of multilevel document clustering and categorization, multidocument summarization using data driven routing to different summarizers and image retrieval linked to text categorization. These research achievements are incorporated into Newsblaster, a deployed prototype which demonstrates the robustness of summarization and TDT technology today. We have begun a large scale online evaluation measuring usage and preferences. Newsblaster is an ongoing project; we are exploring linking summary text phrases directly to the context from which they were drawn, tracking events across days, and incorporating our work on summarizing new information. We are also exploring personalization of Newsblaster, restricting it to user preferred topics or questions. Our vision is the development of system that can provide true updates on current events morning, noon, and night.

Acknowledgements

The work reported here was supported in part by the National Science Foundation under STIMULATE grant IRI-96-18797 and by the Defense Advanced Research Projects Agency under TIDES grant NUU01-00-1-8919. Any opinions, findings, or recommendations are those of the authors and do not necessarily reflect the views of the funding agencies.

8. REFERENCES

[1] R. Barzilay, N. Elhadad, and K. R. McKeown. Sentence ordering in multidocument summarization. In *Proceedings of the 1st Human Language Technology Conference*, San Diego, California, 2001.

[2] R. Barzilay, K. McKeown, and M. Elhadad. Information fusion in the context of multi-document summarization. In *Proc. of the 37th Annual Meeting of the Assoc. of Computational Linguistics*, 1999.

[3] Document understanding conference, 2001.

[4] V. Hatzivassiloglou, L. Gravano, and A. Maganti. An investigation of linguistic features and clustering algorithms for topical document clustering. In *Proceedings of the 23rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-00)*, pages 224–231, Athens, Greece, July 2000.

[5] V. Hatzivassiloglou, J. L. Klavans, and E. Eskin. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 203–212, College Park, Maryland, June 1999. Association for Computational Linguistics.

[6] V. Hatzivassiloglou, J. L. Klavans, M. L. Holcombe, R. Barzilay, M.-Y. Kan, and K. R. McKeown. SIMFINDER: A flexible clustering tool for summarization. In *NAACL Workshop on Automatic Summarization*, pages 41–49. Association for Computational Linguistics, 2001.

[7] K. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, M. Kan, B. Schiffman, and S. Teufel. Columbia multi-document summarization: Approach and evaluation. In *Proceedings of the Document Understanding Workshop (DUC)*, 2001.

[8] K. McKeown, J. Klavans, V. Hatzivassiloglou, R. Barzilay, and E. Eskin. Towards multidocument summarization by reformulatin: Progress and prospects. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence*, 1999.

[9] S. Paek, C. Sable, V. Hatzivassiloglou, A. Jaimes, B. Schiffman, S. Chang, and K. McKeown. Integration of visual and text-based approaches for the content labeling and classification of photographs. In *ACM SIGIR Workshop on Multimedia Indexing and Retrieval (SIGIR-99)*, 1999.

[10] C. Sable and K. W. Church. Using bins to empirically estimate term weights for text categorization. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2001.

[11] C. Sable and V. Hatzivassiloglou. Text-based approaches for non-topical image categorization. *International Journal of Digital Libraries*, **3**(3):261–275, 2000.

[12] B. Schiffman, I. Mani, and K. J. Concepcion. Producing biographical summaries: Combining linguistic knowledge with corpus statistics. In *Proceedings European Association for Computational Linguistics 2001*, 2001.

[13] B. Schiffman, A. Nenkova, and K. McKeown. Experiments in multi-document summarization. In *HLT 2002*, 2002.