Columbia's Newsblaster: New Features and Future Directions

Kathleen McKeown, Regina Barzilay, John Chen, David Elson, David Evans, Judith Klavans, Ani Nenkova, Barry Schiffman and Sergey Sigelman
Department of Computer Science
Columbia University
1214 Amsterdam Avenue, New York, N.Y. 10027
kathy@cs.columbia.edu

Abstract

Columbia's Newsblaster tracking and summarization system is a robust system that clusters news into events, categorizes events into broad topics and summarizes multiple articles on each event. Here we outline our most current work on tracking events over days, producing summaries that update a user on new information about an event, outlining the perspectives of news coming from different countries and clustering and summarizing non-English sources.

1 Introduction

Columbia's Newsblaster[1] provide news updates on a daily basis from news published on the Internet; it crawls news sites, categorizes stories into six broad areas, groups news into stories on the same event, and generates a summary of the multiple articles describing each event. In addition to demonstrating the robustness of current summarization and tracking technology, Newsblaster also serves as a research environment in which we explore new directions and problems. Currently, we are exploring the tasks of multilingual summarization where input sources are drawn from multiple languages and a summary is generated in English on the same event (Figure 1), tracking events across days and generating summaries that update the user on what is new, and editing generated summaries to improve fluency and accuracy. Our focus here is on editing references to people, improving coherency of the summary and ensuring that references are accurate. Editing is particularly important as we add multilingual capabilities, given the errors inherent in machine translation.

[1] http://newsblaster.cs.columbia.edu

2 Multilingual Tracking and Summarization

The multilingual version of Columbia Newsblaster is built upon the English version of Columbia Newsblaster, sharing the same structure and components. To add multilingual capability, the system first crawls web sites in foreign languages, and stores both the language and encoding for the files. To extract the article text from the HTML pages, we use a new article extraction component using language-independent statistical features computed over text blocks along with a machine learning component to classify text blocks as one of "Article Text", "Title", "Image", "Image Caption", or "Other". The article extraction component has been trained and tested on English, Japanese, and Russian data, but is also being successfully applied to French, Spanish, German, and Italian data. We plan to train the article extractor on other languages (Chinese, Arabic, Korean, Spanish, German, French, etc.) in the near future.

To cluster multilingual documents with English documents, we use the existing Newsblaster English document clustering module. Non-English documents are translated for clustering after the article extraction phase. We use simple and fast document translation techniques for clustering if available, since we potentially process thousands of documents for a language for each run. We have developed simple dictionary lookup techniques for translation for clustering for Japanese and Russian; for other languages we use an interface to the Systran translation system via Babelfish. We plan on adding Arabic translation to the system in the near future.

Summarization is performed using the same summarization strategies in Newsblaster. We are experimenting with different methods for improving summary quality when translation of text is noisy. For example, when an input cluster contains both English and foreign sources, we weight the English higher in cases where we determine it is representative of both the English and foreign

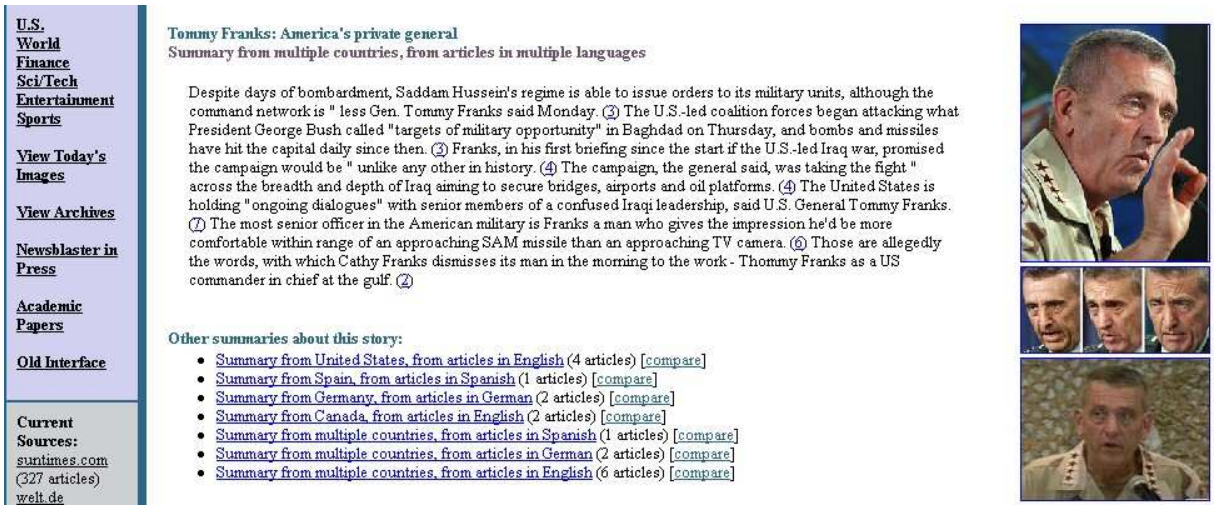

Figure 1: Multilingual Version

input documents. We are also experimenting with methods for determining similarity across documents using different levels of translation.

3 Different Perspectives

When news media report on international issues, they reflect the perspectives of their own countries. In the past, Newsblaster has included all international sources as input to its summaries. Recently, we have added a feature of "international perspectives" to the system. In addition to the universal summary for a particular event, which includes all international sources, Newsblaster now generates separate summaries for each country, which may illustrate unique biases or disagree on facts. The Newsblaster interface allows users to view any pair of summaries side by side to compare different perspectives.

4 Summary Rewrite

Newsblaster also currently includes a module for rewriting summaries to achieve better readability. References to people are rewritten so that the first mention includes the person's full name and a selected description and later mentions are restricted to last name only. In addition to improving readability, the rewritten version of the summary is usually shorter than the version before rewrite, since multiple verbose descriptions of the same entity are discarded. These changes can be seen when comparing the summary sentence with the original document via a link from the summary using a proxy.

5 Event Tracking and Updates

Newsblaster currently identifies events within a single day; a new set of clusters is generated each day. We have designed a new module for tracking events across days, allowing the system to relate stories published on one day to closely related stories on other days. In this way, the user can more easily track events of interest as they unfold. The typical approach for tracking events across days represents each event as one monolithic set of stories. We have focused instead on a model where events on one day can divide into related sub-events on the next day. For example, a set of stories about the start of the Iraq war is an event that can branch into multiple sets of stories, each set representing a different facet of the war. We are currently determining an appropriate evaluation of this approach as well as investigating different possible interfaces.

If a user is tracking events across days, it is more useful to have a summary that provides updates on what is new as opposed to a summary of similarities across all days. We have built a prototype update summarizer that scans new articles extracted by the system and compares these new articles with a background cluster on the same event. The summarizer will provide the user with a summary of only important new developments. As the tracking module locates new articles, it will pass these to the update summarizer, which will determine what, if anything, has changed. This summarizer uses more syntactic and semantic information about the articles to determine novelty than is used in our other summarization strategies and thus, efficiency is a challenge. We will demo these components in a separately from Newsblaster as they have not yet been integrated in the development version.