Chapter 5

Multi-strategy Learning for Topic Detection and Tracking

A joint report of CMU approaches to multilingual TDT

Yiming Yang, Jaime Carbonell, Ralf Brown, John Lafferty, Thomas Pierce, and Thomas Ault

School of Computer Science
Carnegie Mellon University (CMU)
Pittsburgh, PA 15232

Abstract
This chapter reports on CMU's work in all the five TDT-1999 tasks, including segmentation (story boundary identification), topic tracking, topic detection, first story detection, and story-link detection. We have addressed these tasks as supervised or unsupervised classification problems, and applied a variety of statistical learning algorithms to each problem for comparison. For segmentation we used exponential language models and decision trees; for topic tracking we used primarily k-nearest-neighbors classification (also language models, decision trees and a variant of the Rocchio approach); for topic detection we used a combination of incremental clustering and agglomerative hierarchical clustering, and for first story detection and story link detection we used a cosine-similarity based measure. We also studied the effect of combining the output of alternative methods for producing joint classification decisions in topic tracking. We found that a combined use of multiple methods typically improved the classification of new topics when compared to using any single method. We examined our approaches with multi-lingual corpora, including stories in English, Mandarin and Spanish, and multi-media corpora consisting of newswire texts and the results of automated speech recognition for broadcast news sources. The methods worked reasonably well under all of the above conditions.

1. Introduction

Topic Detection and Tracking consists of a set of functionally interrelated tasks, described earlier in this book, and summarized here:
Segmentation: An incoming broadcast news-stream must be segmented into individual, topically-coherent stories. The underlying task is the placement of story boundaries with minimal temporal error.

First-story detection: The onset of newly-breaking news should be signaled, optimally by detecting the first story on a new topic or topic. The underlying task is a pure detection one, minimizing false alarms and misses.

Topic detection: Given all the stories reported in a time window, group the stories into topically-coherent clusters corresponding to individual topics or topics. The underlying task is one of clustering stories, primarily by content similarity but taking into account other factors such as temporal proximity.

Topic tracking: Given one or more news-stories on a given topic or topic, find each future one on that topic or topic as it is reported. The underlying task is supervised learning to induce a classifier based on minimal positive training data (the one or more examples) and much larger sets of negative training data (past history).

Story-link detection: Find all pairs of stories that are linked to each other, primarily by sharing a common topic or topic. The underlying task is one of accurate similarity assessment, but the task could change if different linking criteria were given.

Carnegie Mellon University developed and fielded a variety of methods for these tasks, as described in sections below corresponding to each task. This chapter focuses more on recent CMU results, especially on multi-classifier tracking and story-link detection; previous results have been reported in the literature[21][20]. In general, each of the three years of TDT research witnessed progress in terms of increasing sophistication of methods used, more challenging data sets, and better or more robust overall results.

Each year of active TDT research produced new training and testing corpora, starting from TDT pilot study[1] (which consisted of monolingual newswire data) and continuing through TDT-1998 and TDT-1999 (which consisted of English and Mandarin data from newswire and broadcast news sources). Moreover, CMU labeled Spanish data collected contemporaneously with TDT-1998 and TDT-1999, but not part of the "official" TDT program, and we report on results obtained from that data as well. Although the topics labeled in the Spanish data are the same ones as in the English and Mandarin, the reporting periods do not coincide, making exact comparisons difficult. Nonetheless, Spanish results, as reported below, are in the equivalent-to-better range, compared to English and Mandarin ones. Each set of results reported herein is labeled according to the data set(s) used, although no additional system tuning was done for Spanish.