Domain-informed Topic Detection

Cormac Flynn and John Dunnion

Department of Computer Science,
University College Dublin,
Ireland.
{cormac.flynn,john.dunnion}@ucd.ie

Abstract. We discuss *Topic Detection*, a sub-task of the *Topic Detection and Tracking (TDT)* Project, and present a system that uses the linguistic and temporal features of news reportage to enhance the discovery of events in a collection of news articles. We describe an online application of these techniques that constructs topical clusters from live news feeds. We conclude that these approaches promise more coherent and useful clusters and suggest some areas of future work.

1 Introduction

The *Topic Detection and Tracking (TDT)* Project [1] is a DARPA-sponsored initiative [1] that aims to develop systems that monitor news-wire, broadcast news and other topical information sources, detect breaking stories and new events, and track these as they change over time.

The most recent phase of the project defined five tasks that are necessary for an effective TDT system: Story Segmentation, Topic Tracking, Topic Detection, New Event Detection and Link Detection. We concentrate on *Topic Detection*, i.e. the grouping together of related documents into topically cohesive clusters that correspond to distinct news events. This can be performed on a retrospective collection of documents or on a live stream. Such a system could be used to group an accumulation of time-ordered news articles into topical clusters, or to monitor the many online sources of news reportage.

This paper consists of six sections. We describe our basic Topic Detection system in Section 2. Section 3 discusses those features that distinguish problems in the TDT domain from general information filtering. Section 4 describes extensions to the basic system that exploit these features. Section 5 presents a Topic Detection system that operates on live news feeds. Finally, we give our conclusions and suggest future work in Section 6.

2 Baseline Topic Detection System

Our baseline Topic Detection system accepts documents from a variety of sources; from the TDT corpora, the TREC collection, or from online RSS feeds. For the

[1] http://www.nist.gov/speech/tests/tdt/