Lexical Chains versus Keywords for Topic Tracking

Joe Carthy

Department of Computer Science,
University College Dublin, Ireland.
Joe.Carthy@ucd.ie

Abstract. This paper describes research into the use of lexical chains to build effective Topic Tracking systems and compares the performance with a simple keyword-based approach. Lexical chaining is a method of grouping lexically related terms into so called *lexical chains*, using simple natural language processing techniques. Topic tracking involves tracking a given news event in a stream of news stories i.e. finding all subsequent stories in the news stream that discuss the given event. This paper describes the results of a novel topic tracking system, LexTrack, based on lexical chaining and compares it to a keyword-based system designed using traditional IR techniques.

1 Introduction

Topic detection and tracking research has grown out of a DARPA-sponsored initiative to investigate the computational task of finding new events and tracking existing events in a stream of textual news stories from multiple sources [1]. These sources include news broadcast programs such as CNN news and newswire sources such as Reuters. The information in these sources is divided into a sequence of stories that provide information on one or more events. The tracking task is defined as that of associating incoming stories with events known to the system. An event is defined as "known" by its association with stories that discuss the event. So, each target event is defined by a list of stories that define it. If we take an event such *as "the Kobe earthquake",* then the first story (or first N stories) in the corpus describing the Kobe earthquake could be used as the definition of that event.

A TDT test corpus was constructed to facilitate the TDT initiative. This corpus includes 15,863 news stories from July 1, 1994 to June 30, 1995. The corpus included relevance judgments for a set of 25 events covering a broad spectrum of interests such as disaster stories (e.g. Kobe earthquake in Japan) and crime stories (e.g. OJ Simpson trial). Every story in the corpus was judged with respect to every event by two sets of assessors and any conflicts were reconciled by a third assessor.

2 Lexical Chaining

The notion of lexical chaining derives from work in the area of textual cohesion by Halliday and Hasan [2]. The linguistics term *text* is used to refer to any passage spoken or written that forms a unified whole. This unity or cohesion may be due, for