Topic Tracking Based on Linguistic Features

Fumiyo Fukumoto and Yusuke Yamaji

Interdisciplinary Graduate School of Medicine and Engineering,
Univ. of Yamanashi, 4-3-11, Takeda, Kofu, 400-8511, Japan
fukumoto@yamanashi.ac.jp, g03mk031@ccn.yamanashi.ac.jp

Abstract. This paper explores two linguistically motivated restrictions on the set of words used for topic tracking on newspaper articles: named entities and headline words. We assume that named entities is one of the linguistic features for topic tracking, since both topic and event are related to a specific place and time in a story. The basic idea to use headline words for the tracking task is that headline is a compact representation of the original story, which helps people to quickly understand the most important information contained in a story. Headline words are automatically generated using headline generation technique. The method was tested on the Mainichi Shimbun Newspaper in Japanese, and the results of topic tracking show that the system works well even for a small number of positive training data.

1 Introduction

With the exponential growth of information on the Internet, it is becoming increasingly difficult to find and organize relevant materials. Tracking task, i.e. starts from a few sample stories and finds all subsequent stories that discuss the target topic, is a new line of research to attack the problem. One of the major problems in the tracking task is how to make a clear distinction between a topic and an event in the story. Here, an event refers to the subject of a story itself, i.e. a writer wants to express, in other words, notions of who, what, where, when, why and how in the story. On the other hand, a topic is some unique thing that occurs at a specific place and time associated with some specific actions. It becomes background among stories. Therefore, an event drifts, but a topic does not. For example, in the stories of ‘Kobe Japan quake’ from the TDT1 corpus, the event includes early reports of damage, location and nature of quake, rescue efforts, consequences of the quake, and on-site reports, while the topic is Kobe Japan quake.

A wide range of statistical and machine learning techniques have been applied to topic tracking, including k-Nearest Neighbor classification, Decision Tree induction, relevance feedback method of IR, hierarchical and non-hierarchical clustering algorithms, and a variety of Language Modeling. The main task of these techniques is to tune the parameters or the threshold for binary decisions to produce optimal results. In the TDT context, however, parameter tuning is a tricky issue for tracking. Because only the small number of labeled positive stories is available for training. Moreover, the well-known past experience from IR that notions of who, what, where, when, why, and how may not make a great contribution to the topic tracking task causes this fact, i.e. a topic and an event are different from each other.

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 10–21, 2005.
© Springer-Verlag Berlin Heidelberg 2005
This paper explores two linguistically motivated restrictions on the set of words used for topic tracking on newspaper articles: named entities and headline words. A topic is related to a specific place and time, and an event refers to notions of who (person), where (place), when (time) including what, why and how in a story. Therefore, we can assume that named entities is one of the linguistic features for topic tracking. Another linguistic feature is a set of headline words. The basic idea to use headline words for topic tracking is that headline is a compact representation of the original story, which helps people to quickly understand the most important information contained in a story, and therefore, it may include words to understand what the story is about, what is characteristic of this story with respect to other stories, and hopefully include words related to both topic and event in the story. A set of headline words is automatically generated. To do this, we use a technique proposed by Banko [2]. It produces coherent summaries by building statistical models for content selection and surface realization. Another purpose of this work is to create Japanese corpus for topic tracking task. We used Mainichi Shimbun Japanese Newspaper corpus from Oct. to Dec. of 1998 which corresponds to the TDT3 corpus. We annotated these articles against the 60 topics which are defined by the TDT3.

The rest of the paper is organized as follows. The next section provides an overview of existing topic tracking techniques. We then describe a brief explanation of a headline generation technique proposed by Banko et al. [2]. Next, we present our method for topic tracking, and finally, we report some experiments using the Japanese newspaper articles with a discussion of evaluation.

2 Related Work

The approach that relies mainly on corpus statistics is widely studied in the topic tracking task, and an increasing number of machine learning techniques have been applied to the task. CMU proposed two methods: a k-Nearest Neighbor (kNN) classifier and a Decision-Tree Induction (dtree) classifier [1, 20, 3]. Dragon Systems proposed two tracking systems; one is based on standard language modeling technique, i.e. unigram statistics to measure story similarity [18] and another is based on a Beta-Binomial model [10]. UMass viewed the tracking problem as an instance of on-line document classification, i.e. it classifies documents into categories or classes [4, 8, 19, 9, 14]. They proposed a method including query expansion with multi-word features and weight-learning steps for building linear text classifiers for the tracking task [13]. These approaches, described above, seem to be robust and have shown satisfactory performance in stories from different corpora, i.e. TDT1 and TDT2. However, Carbonell claims that something more is needed if the system is intended for recognizing topic drift [3]. Yang et al. addressed the issue of difference between early and later stories related to the target event in the TDT tracking task. They adapted several machine learning techniques, including k-Nearest Neighbor(kNN) algorithm and Rocchio approach [21]. Their method combines the output of a diverse set of classifiers and tuning parameters for the combined system on a retrospective corpus. The idea comes from the well-known practice in information retrieval and speech recognition of combining the output of a large number of systems to yield a better result than the individual system’s output. They reported that the new variants of kNN reduced up to 71% in weighted error rates on the TDT3-dryrun corpus.