Generating Headline Summary from a Document Set

Kamal Sarkar[1], Sivaji Bandyopadhyay[1]

[1] Computer Science & Engineering Department, Jadavpur University,
Kolkata – 700 032, INDIA
jukamal2001@yahoo.com, sivaji_ju@vsnl.com

Abstract. This paper discusses an approach to generate headline summary from a set of documents. Headline summary is basically a very short summary in the form of headline. As the amount of on-line information increases, systems that can automatically summarize multiple documents are becoming increasingly desirable. In this situation, headline summary is useful for users who only need information on the main topics in a set of documents. Headline summary from multiple documents will be very useful in the text mining applications for the generation of meaningful label (a compact identifier that allows a person to quickly see what the topic is about) for a cluster of documents.

1 Introduction

In this paper we present a system that will cluster the text documents collected from multiple online sources and generate a headline summary in one or two sentences for each cluster by identifying named entities from each document in the set to make a global list of named entities and forming a headline summary for the set.

All the previous work on headline generation [1, 2] was done on single documents but the focus in the present work is on headline summary generation from a set of documents. Moreover, instead of using only statistical approaches, we have used named entity cues and summary generation techniques for our work, since it is very difficult to have a training corpus of document set—headline pairs. In the next section we present the proposed approach. The system is evaluated in Section 3.

2 Proposed Approach

News collected from multiple sources should be clustered. Clustering technique adopted is similar to the method used by Chen and Lin [3].

The input to our system is a cluster of related documents. Based on the observations of human-produced headline summary, we have developed the following algorithm.