Proceedings of the 2013 Federated Conference on
Computer Science and Information Systems pp. 577–584

Automatic Identification of Broadcast News Story Boundaries Using the Unification Method for Popular Nouns

Zainab Ali Khalaf[1,2]
[2]Department of Computer Science, College of Science, University of Basra, Iraq
Email: zainab_ali2004@yahoo.com

Tan Tien Ping
[1]School of Computer Sciences, Universiti Sains Malaysia USM, 11800 Penang, Malaysia
Email: tienping@cs.usm.my

Abstract—Herein we describe the latent semantic algorithm method for identifying broadcast news story boundaries. The proposed system uses the pronounced forms of words to identify story boundaries based on popular noun unification. Commonly used clustering methods use latent semantic analysis (LSA) because of its excellent performance and because it is based on deep semantic rather than shallow principles. In this study, the LSA algorithm with and without unification was used to identify boundaries of Malay spoken broadcast news stories. The LSA algorithm with the noun unification approach resulted in less error and better performance than the LSA algorithm without noun unification.

Keywords: spoken document; broadcast news; story boundary identification; latent semantic analysis

I. Introduction

Because nouns bear more semantic meaning than other parts of speech and because they are the main characteristics used to identify documents stories [1], natural language processing applications often focus on nouns as essential components of the documents being processed. Names of persons, for example, are useful noun components in natural language processing, especially during automatic sentence clustering. In recent years, spoken document processing has become a popular and interesting topic within the field of natural language processing. In general, spoken document processing adapts natural language processing applications using speech input rather than text input.

Processing spoken documents is challenging because of the word errors generated by the automatic speech recognition (ASR) process [2], [3]. Determining the boundaries of broadcast news stories is another obstacle to processing spoken documents. The lack of overt punctuation and formatting contributes to this problem. In order to retrieve information, the beginning and the end of the segments or paragraphs within a document must be determined [3]-[6]. The process of determining the boundaries of the segments in the text is not an easy process [3]-[7].

Word errors generated by the ASR process can occur when recordings are made in a noisy environment or when pronunciation is unclear. The latter is especially true for vowel letters. An example from a Malay broadcast news story is as follows: The name of a professional badminton player was written four different ways in four sentences when converted from spoken news to written news by the ASR system (lee chong wei, choong wei, chong wee, and chan wee). The conversion problem was related to the vowel sound, in that the [u:] sound can be written as “oo, o, ou, ew, ue, u, and ui” and the [i:] sound can be written as “ee, ea, ei, and ie.” Silent sounds (pronounced n+ unpronounced g) also pose problems for ASR [8], [9].

Identification of story boundaries with the added problem of pronunciation errors is a complicated task. It requires human knowledge of the rules of correct pronunciation of lexical items. To address these problems, we propose a new method to improve story boundary identification in spoken documents using the popular noun unification approach.

II. Related Work

The absence of punctuation and capitalization in spoken documents makes it challenging to automatically identify story boundaries in multimedia documents. Previous attempts have concentrated on three types of cues: visual cues, such as the presence of an anchor’s face [7] or motion changes [7]; audio cues, such as significant pauses or reset of pitch; and lexical cues, such as word similarity measures within speech recognition transcripts or closed captions of video [10], [11]. Cues from completely different modalities (audio, video, and text) are often consolidated to achieve better story boundary identification [7], [12].

Hearst et al. (1997) proposed the TextTiling approach to story boundary identification [10]. It is based on the straightforward observation that different topics usually employ different sets of words and that shifts in vocabulary usage are indicative of topic changes [10]. As a result, pairwise

978-1-4673-4471-5/$25.00