ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language xxx (2014) xxx–xxx

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Topic segmentation of TV-streams by watershed transform and vectorization[☆]

Vincent Claveau [a,*], Sébastien Lefèvre [b,1]

Q1

[a] IRISA – CNRS, Campus de Beaulieu, F-35042 Rennes, France
[b] IRISA – Univ. Bretagne-Sud, Campus de Tohannic, F-56017 Vannes, France

Received 18 April 2013; received in revised form 18 January 2014; accepted 28 April 2014

Abstract

A fine-grained segmentation of Radio or TV broadcasts is an essential step for most multimedia processing tasks. Applying segmentation algorithms to the speech transcripts seems straightforward. Yet, most of these algorithms are not suited when dealing with short segments or noisy data. In this paper, we present a new segmentation technique inspired from the image analysis field and relying on a new way to compute similarities between candidate segments called Vectorization. Vectorization makes it possible to match text segments that do not share common words; this property is shown to be particularly useful when dealing with transcripts in which transcription errors and short segments makes the segmentation difficult. This new topic segmentation technique is evaluated on two corpora of transcripts from French TV broadcasts on which it largely outperforms other existing approaches from the state-of-the-art.

Keywords: Watershed transform; Image segmentation; Vectorization; Topic segmentation

1. Introduction

Topic segmentation is of high interest in Multimedia information retrieval. Indeed, it is needed to perform automatic structuring of TV streams, a keystone for every processing of such streams, which is still done manually in national archive agencies like the French INA. A way to obtain this structuration is to first transcribe the audio tracks of the TV streams into textual data, and then perform the topic segmentation from textual data to split the streams into semantic units (e.g., reports).

In this paper we address the problem of topic segmentation of speech in this applicative framework based on a twofold contribution.[2] First, our topic segmentation system is based on the watershed paradigm derived from image segmentation. Second, a key component for this approach is the calculation of the similarity between two successive

[☆] This paper has been recommended for acceptance by E. Briscoe.

Q2 [*] Corresponding author. Tel: +33 0299847447; fax: +33 0299847171.
E-mail addresses: vincent.claveau@irisa.fr (V. Claveau), sebastien.levevre@irisa.fr (S. Lefèvre).
[1] Tel: +33 0297017266; fax: +33 0297017279.

[2] Preliminary versions and results of this system have been published in Claveau and Lefèvre (2011).

http://dx.doi.org/10.1016/j.csl.2014.04.006

possible segments; in this paper we present a new technique, called *vectorization* that we recently introduced in the information retrieval field.

The paper is organized as follows. We first present state-of-the-art approaches used for topic segmentation. We then show that topic segmentation and image segmentation have common characteristics (Section 3). From this observation we build a topic segmentation method based on the watershed transform, a common morphological tool that identifies segments or regions within a topographic surface. We suggest to build this topographic surface with the help of vectorization which we think is especially suited when dealing with small segments or noisy data such as TV streams (Section 4). A first set of experiments, whose goal is to assess the performance of this approach on a standard segmentation benchmark, is presented in Section 5. Then, experiments performed on two real TV broadcast corpora are presented and discussed (Section 6). Finally, Section 7 concludes this work and provides future research directions.

2. Related work

This section is divided into two parts. The first one presents state-of-the-art techniques for topic segmentation. In the second subsection, we compare those techniques and image segmentation ones and show that both fields share many similarities that explain our choice of the watershed transform as a basis for our approach.

2.1. Approaches for topic segmentation

Topic segmentation in TV streams has addressed in several ways in the literature. Multimodal approaches have been proposed, which makes the most of audio, visual and speech features. For instance, segmentation of TV news reports have been explored in TRECVID competition.[3] (Amir et al., 2004, for a representative multimodal system) It is worth noting that most of the approaches proposed chiefly rely on training data and do not extend well to other dataset as they focus on superficial clues such as anchor person recognition or background colors. While Poulisse and Moens (2009) have shown the interest of adding multimodal features to improve text-based story segmentation, in the remaining, we only focus on text only approaches.

Various approaches have been applied to speech or text based topic segmentation. Several methods rely on some particularities of the document format, on the detection discourse markers either given by experts (Christensen et al., 2005), or automatically learned (Beeferman et al., 1999). Such techniques require well-formed text and especially a grammatically correct sentence tokenization; they are therefore not suited for texts generated from automatic speech recognition (ASR) systems in which the concept of sentence can rarely match with the oral specifics. Conversely, another kind of approaches is to detect topic changes through document content analysis. These content-based approaches yield high performances and they are less dependent to the document formatting. The overall good quality of modern ASR systems (Ostendorf et al., 2008) makes the use of such approaches on transcribed texts possible (Mulbregta et al., 1999). This is also the approach adopted in our system. In the following, we present representative content-based techniques from the state-of-the-art.

The segmentation process of SEGMENTER (Kan et al., 1998) relies on a representation of the text as weighted lexical chains. Finding the boundaries is thus equivalent to partitioning the resulting graph. The two approaches in DOTPLOTTING (Reynar, 2000) and C99 (Choi, 2000) differ in the way the content is represented, but both rely on the computation of similarities between the candidate segments and then on a clustering based on the resulting similarity matrix. Utiyama and Isahara (2001) propose to use a statistical approach based on hidden Markov models. Here again, the lexical cohesion, key component of the approach, is measured classically with the help of language modeling. The computation of similarity is also at the heart of the TEXT-TILING system (Hearst, 1997), in which a sliding window is used to compare the content before and after each possible boundary. The similarity measure used is inspired from the information retrieval domain (for instance, a cosine computed from TF or TF-IDF vector representation of the context), and the final boundaries are searched among the places in which the lexical cohesion reaches a significant local minimum. It is worth noting that the approach proposed in this paper can be seen as an modern version of TEXT-TILING in which the similarity computation is done through vectorization (see Section 4) and in which the watershed extends the simple boundary detection process used in the original version of TEXT-TILING (cf. next subsection).

[3] http://www-nlpir.nist.gov/projects/tv2004/tv2004.html.

Please cite this article in press as: Claveau, V., Lefèvre, S., Topic segmentation of TV-streams by watershed transform and vectorization. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.04.006