Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 37 (2016) 24–46

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Weighted hierarchical archetypal analysis for multi-document summarization[☆]

Ercan Canhasi [*], Igor Kononenko

Laboratory for Cognitive Modeling, Faculty of Computer and Information Science, University of Ljubljana, Tržaška cesta 25, 1000 Ljubljana, Slovenia

Received 22 July 2014; received in revised form 14 November 2015; accepted 17 November 2015
Available online 23 November 2015

Abstract

Multi-document summarization (MDS) is becoming a crucial task in natural language processing. MDS targets to condense the most important information from a set of documents to produce a brief summary. Most existing extractive multi-document summarization methods employ different sentence selection approaches to obtain the summary as a subset of sentences from the given document set. The ability of the weighted hierarchical archetypal analysis to select "the best of the best" summary sentences motivates us to use this method in our solution to multi-document summarization tasks. In this paper, we propose a new framework for various multi-document summarization tasks based on weighted hierarchical archetypal analysis. The paper demonstrates how four variant summarization tasks, including general, query-focused, update, and comparative summarization, can be modeled as different versions acquired from the proposed framework. Experiments on summarization data sets (DUC04-07, TAC08) are conducted to demonstrate the efficiency and effectiveness of our framework for all four kinds of the multi-document summarization tasks.

Keywords: Multi-document summarization framework; Weighted hierarchical archetypal analysis; General; Query-focused; Update; Comparative summarization

Multi-document summarization is an integral tool for document understanding. MDS enables better information benefits by creating brief and descriptive summaries for a large collection of documents. It has been found various uses in many real world applications. For example, multi-document summarization can be applied to summarize user-interested news events from a huge pool of news. A vast amount of available online texts are responsible for serious difficulties in development of the question answering or information retrieval systems. Fortunately, by extracting the vital information from documents, multi-document summarization can make possible development of the question answering or information retrieval systems. Note that the generated summary can be: (a) the general, non-restricted or typical (Mani, 2001); (b) the query-focused, most typical or archetypal summary which is explicitly biased toward a user defined query; (c) the comparative summary (Wang et al., 2012) which recaps dissimilarities between corresponding document groups; (d) and the update summary (Dang and Owczarzak, 2008) which produces very brief extract of the latest documents to apprehend novel information distinct from earlier documents.

[☆] This paper has been recommended for acceptance by Shrikanth Narayanan.

[*] Corresponding author. Tel.: +386 37745542501.
E-mail address: ercan.canhasi@uni-prizren.com (E. Canhasi).

http://dx.doi.org/10.1016/j.csl.2015.11.004
0885-2308/© 2015 Elsevier Ltd. All rights reserved.

In this paper, we propose a new framework for MDS using the weighted hierarchical archetypal analysis (wHAA-Sum). Four different and known summarization tasks namely general, query-focused, update, and comparative summarization tasks, can be modeled as different versions acquired from the proposed framework. An effective foundation to promote their differences while settling the affinities among different summarization tasks is served by this framework.

Generally, the main topic of a document set is composed of some sub-topics. The assumption is that capturing the sub-topic structure is essential for a successful MDS system. These sub-topics can be treated using cluster-based methods by modeling the logic structure of the topics and sub-topics. However, the implicit structure of the topic can not only be represented by the explicit features such as statistical term distributions. In this paper, we argue that sentence selection in a MDS can be based on the implicit structure of the topic covered in the document set. By representing the relationship information with graph, we research the use of sub-topics as a model of the document collection, where sub-topics are represented as sub-archetypes. In this way, we investigated ways of selecting the most salient sentences from important sub-topic originating from significant topics by utilizing the hierarchical archetypal analysis.

In our summarization framework, the multi-document summarization problem is firstly generalized to the weighted hierarchical archetypal analysis problem. Then several useful properties of the wHAA are identified and taken into consideration for the greedy summarization algorithm. The latter is further shown to have the ability of addressing the multi-document summarization problem. We finally use this algorithm to propose the framework for different multi-document summarization tasks.

Our work displays benefits from two perspectives:

1. It proposes a new generic framework to address different summarization problems.
2. It proposes a novel version of the well-known archetypal analysis algorithm, namely the weighted hierarchical archetypal analysis algorithm.

The rest of the paper is organized as follows. In Section 1, we review the related work about different multi-document summarization tasks and the submodular function. After introducing the original archetypal analysis (AA) and weighted archetypal analysis (wAA) algorithms and after proposing the novel hierarchical version of wAA in Section 2, we propose the hierarchical wAA based summarization method in Section 3. Section 4 presents the framework for multidocument summarization, and shows how to model the four aforementioned summarization tasks. Section 5 presents experimental results of our framework on well accepted summarization data sets. Finally, Section 6 concludes the paper.

1. Related work

1.1. Document summarization

In a broad sense there are two types of summarization schemes: extractive and abstractive. Extractive summarization reduces the problem of summarization into the problem of selecting a most significant subset of the sentences in the original document set. Abstractive summarization is harder since it requires the composition of novel sentences, unseen in the original sources.

In the non-restricted, general summarization each sentence is scored with a significance value, and then sentences are ranked based on the significance score. The significance scores are commonly calculated as a combination of syntactic, semantic and statistical characteristics. The well-known baselines for extractive multi-document summarization can be categorized into one of the following general models: Centrality-based (Radev et al., 2004; Ribeiro and de Matos, 2011; Erkan and Radev, 2004), maximal marginal relevance (MMR) (Carbonell and Goldstein, 1998; Sanner et al., 2011), coverage-base methods (Lin and Hovy, 2000; Sipos et al., 2012), and hybrids (centrality + coverage-based) (Marujo et al., 2015; Ribeiro et al., 2013) Centrality-as-relevance methods base the detection of the most salient passages on the identification of the central passages of the input source(s). MMR methods are based on a measure for quantifying the extent of dissimilarity between the sentences being considered and those already selected. Coverage-based summarization defines a set of concepts that need to occur in the sentences selected for the summaries. Although wHAASum is not a typical example of any of this groups, it mostly belongs to centrality based methods.