Neurocomputing 000 (2018) 1–12

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Unsupervised measure of Chinese lexical semantic similarity using correlated graph model for news story segmentation

Wei Feng[a,b], Xuecheng Nie[a,b,c,*], Yujun Zhang[a,b], Lei Xie[d], Jianwu Dang[a,e]

[a] School of Computer Science and Technology, Tianjin University, Tianjin 300350, China
[b] Key Research Center for Surface Monitoring and Analysis of Cultural Relics, State Administration of Cultural Heritage, China
[c] The Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[d] School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China
[e] The Department of Electrical and Computer Engineering, School of Information Science, JAIST, Japan

ARTICLE INFO

Article history:
Received 24 January 2018
Revised 25 June 2018
Accepted 16 August 2018
Available online xxx

Communicated by T. Mu

Keywords:
Story segmentation
Unsupervised correlated affinity graph (UCAG) model
Contextual correlation
Common character correlation
Parallel affinity propagation
Generalized cosine similarity

ABSTRACT

This paper presents a simple yet effective approach to unsupervisedly measuring Chinese lexical semantic similarity, and shows its promising performance in automatic story segmentation of Mandarin broadcast news. Our approach centers on the unsupervised correlated affinity graph (UCAG) model, which is initialized as a hybrid sparse graph, encoding both explicit word-to-word contextual correlations and latent word-to-character correlations within the given corpus. The UCAG model further diffuses the initial sparse correlations throughout the graph by parallel affinity propagation. This provides us with a dense, reliable, and corpus-specific lexical semantic similarity measure, which comes from purely unlabeled data. We then generalize the classical cosine similarity metric to effectively take soft similarities into account for story segmentation. Extensive experiments on benchmark datasets validate the superiority of the proposed similarity measure over previous measures. We specifically show that our similarity measure averagely helps to achieve 7.7% relative F1-score improvement to the accuracy of state-of-art normalized cuts (NCuts) based story segmentation on two holistic benchmark Mandarin broadcast news corpora, TDT2 and CCTV, and achieves 10.8% relative F1-score improvement on the detailed broadcast news subsets.

1. Introduction

Story segmentation aims at partitioning a text, audio or video stream into a sequence of topically coherent segments, known as stories. As the explosive growth of multimedia content on the web, there is an urgent demand for reliable automatic story segmentation techniques. This is because the well-segmented stories are important prerequisite for various content-based tasks to organize the massive amount of multimedia data, such as topic tracking [1], summarization [2], information extraction, indexing and retrieval [3].

Previous studies on automatic story segmentation mainly focus on the exploration of topic boundary cues and segmentation criteria. Topic boundary cues can be generally divided into three categories: audio cues, such as pause and pitch [4]; video cues, such as anchor face and correlation between frames [5]; and lexical cues from text transcripts [6]. Among these cues, lexical cues are of great interest, since they reveal topic shift via semantic variation based on the words usage, which is independent of various editorial rules of broadcast news.

Based on lexical cues, various segmentation criteria have been proposed to detect topic shifts, e.g. TextTiling [7], latent Dirichlet allocation (LDA) [8], maximum lexical cohesion (MLC) [9]. These lexical-based story segmentation criteria work well for news transcripts with sharp topic shifts in lexical distribution. However, the lexical transitions between different topics may be smooth in real-world broadcast news data. Recently, graph-theoretic methods have shown promising potentials in automatic story segmentation for processing diverse topic lexical shifts. Through graph embedding, sentences are represented as nodes, and the pairwise relations between sentences are represented as edges. Then, story segmentation transfers to a graph partitioning problem, of which min-

* Corresponding author at: Vision & Machine Learning Lab, Block E4 #08-24, 4 Engineering Drive 3, The Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583.
E-mail addresses: wfeng@tju.edu.cn (W. Feng), niexuecheng@u.nus.edu (X. Nie), yujunzhang@tju.edu.cn (Y. Zhang), lxie@nwpu-aslp.org (L. Xie), dangjianwu@tju.edu.cn (J. Dang).

https://doi.org/10.1016/j.neucom.2018.08.061
0925-2312/

Table 1
The accuracy comparison of story segmentation using corpus driven based methods and general knowledge based methods for word-level semantic similarity measure on TDT2 dataset.

Corpus driven based method				General knowledge based method
UCAG	PMI	LSA	pLSA	HowNet
0.7212	0.7127	0.6958	0.6964	0.6949

imum normalized cuts (NCuts) is one of the most successful criteria in story segmentation and document analysis [10–13].

Despite the notable recent successes in lexical story segmentation via graph theoretic models, there is a major drawback of existing lexical-based story segmentation methods, that is, they usually ignore the latent word-level semantic similarities [14], which, in contrast to image segmentation [15,16], are very important in locating the topic shifts, thus may highly affect the segmentation performance. In contrast to the widely studied story segmentation criteria, most of the existing methods just simply employ the repetition-based hard lexical similarity $S_H(w_a, w_b)$ between the two words w_a and w_b, which is defined as:

$$S_H(w_a, w_b) = \begin{cases} 1 & \text{if } a = b, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

From Eq. (1), we can observe that the widely-used hard lexical similarity only considers the equivalence of the same words, while the latent semantic correlations between different words are simply ignored. This might help to maintain the simplicity and efficiency of the algorithms, and make the story segmentation accuracy mainly rely on the partitioning criteria. However, it is clear that properly taking the latent semantic correlations between different words into consideration is much more desirable to improve lexical similarity measurements, thus certainly helps to boost the segmentation performance.

Previous studies on measuring lexical semantic similarity can be categorized into two classes: *i.e.* corpus driven based methods and general knowledge based methods. General knowledge based methods, such as WordNet [17] and DISCO [18], are derived from POS-tagged corpus with conscientious labeling of linguistic experts according to the conceptual and semantic senses of nouns, verbs, adjectives and adverbs, respectively. These measures are usually linguistically meaningful and generally corpus-independent, thus can be used for the pre-computed word spaces. Currently, WordNet and DISCO have supported English, French, German, etc. However, they are not yet mature for Chinese language. Recently, Liu and Li [19] proposed a general knowledge based method for measuring Chinese word-level semantic similarity using HowNet [20]. Despite the success of word-level semantic similarity derived from general knowledge based approaches, they are not the best choice for Chinese story segmentation because these approaches are relied on the knowledge of linguistic experts and the semantic similarities among words are always annotated by the subjective opinion, which is not suitable for story segmentation.

A lot of corpus driven based approaches have been proposed in recent years, such as Pointwise Mutual Information (PMI) [21], Latent Semantic Analysis (LSA) [22], Probabilistic Latent Semantic Analysis (pLSA) [23]. These corpus driven based approaches utilize statistical or mathematical analysis of the word distribution to generate semantic similarity related to the given corpus. We have conducted an experiment to compare the story segmentation accuracy using the highest 200 word-level semantic similarities generated from corpus driven based and general knowledge based methods, respectively, where UCAG denotes the proposed semantic similarity measure in this paper. And the results are listed in Table 1. From Table 1, we can find that semantic similarities generated from corpus driven based methods can always outperform general knowledge based methods. The experimental results can illustrate that word-level semantic similarity generated form corpus driven based are much preferred than general knowledge based approaches in story segmentation.

In this paper, we propose a corpus driven based method for measuring Chinese lexical similarity by unsupervised correlated affinity graph, which is utilized for NCuts based story segmentation. Unlike most of existing corpus driven based semantic similarity measurements, our approach takes explicit word-to-word contextual correlations into account to reflect the lexical correlations between Chinese words. In addition, we find that latent word-to-character correlations can help to produce more reasonable semantic similarity measurement. By combining the explicit word-to-word contextual correlations and latent word-to-character correlations together, we construct an unsupervised correlated affinity graph that embodies the Chinese lexical semantic similarities. Moreover, a corpus-dependent semantic similarity measurement is derived through parallel affinity propagation. We then present an extended cosine similarity metric to encode the derived semantic similarity. Experiments on benchmark datasets CCTV and TDT2 show that, story segmentation using proposed semantic similarity measurement can achieve superior performance over state-of-the-art semantic similarity measurements and other story segmentation approaches.

In the following, we first briefly review the existing methods of story segmentation and semantic similarity measurement in section 2. In Section 3, we elaborate our unsupervised correlated graph model and the semantic similarity measurement in detail. Experimental results on benchmark datasets are shown in Section 4, followed by conclusions in Section 5.

2. Related work

2.1. Automatic story segmentation

Previous lexical story segmentation approaches can be divided into three main categories. The first category is based on topic modeling, e.g., Latent Dirichlet Allocation (LDA) [8]. They treat word sequences as observations of some latent topics. By some optimal criteria, a sequence of topic labels is assigned to the input text or speech transcript. Then the segmentation is obtained simply by marking boundaries between every pair of adjacent parts with different topic labels. In contrast, the second category directly investigates word usage and segments the input stream into lexically cohesive parts. A typical method is TextTiling [7]. Based on an intuitive idea that different topics usually employ different sets of words, TextTiling scans the text and marks a boundary when lexical similarity of two adjacent sentences is lower than a tuned threshold. The third category [9,10] aims at finding an optimal segmentation under some global criteria, rather than merely detecting local shifts. For instance, Malioutov et al. [10] formulated story segmentation as a sentence-level graph partitioning problem by optimizing the normalized cuts (NCuts) criterion. Liu et al. [9] proposed a lexical cohesion approach to news story segmentation, which measures the lexical cohesion of a segment by KL-divergence from its word distribution to an associated piecewise uniform distribution and is able to detect story boundaries at finer word/subword granularity. All of these methods focused on what criterion or model to be used in story segmentation, but the lexical semantic similarity measurement is ignored, which can help to improve the performance of automatic story segmentation.

Please cite this article as: W. Feng et al., Unsupervised measure of Chinese lexical semantic similarity using correlated graph model for news story segmentation, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.08.061