Knowledge-Based Systems xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Exploring events and distributed representations of text in multi-document summarization

Q1 Luís Marujo [a,c,d,*], Wang Ling [a,c,d], Ricardo Ribeiro [a,b], Anatole Gershman [d], Jaime Carbonell [d], David Martins de Matos [a,c], João P. Neto [a,c]

[a] INESC-ID Lisboa, Rua Alves Redol, 9, 1000-029 Lisboa, Portugal
[b] ISCTE-Instituto Universitário de Lisboa, Av. das Forças Armadas, 1649-026 Lisboa, Portugal
[c] Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal
[d] Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, 15213-3891, USA

A R T I C L E I N F O

Article history:
Received 17 July 2015
Revised 25 September 2015
Accepted 5 November 2015
Available online xxx

Keywords:
Multi-document summarization
Extractive summarization
Event detection
Distributed representations of text

A B S T R A C T

In this article, we explore an event detection framework to improve multi-document summarization. Our approach is based on a two-stage single-document method that extracts a collection of key phrases, which are then used in a centrality-as-relevance passage retrieval model. We explore how to adapt this single-document method for multi-document summarization methods that are able to use event information. The event detection method is based on Fuzzy Fingerprint, which is a supervised method trained on documents with annotated event tags. To cope with the possible usage of different terms to describe the same event, we explore distributed representations of text in the form of word embeddings, which contributed to improve the summarization results. The proposed summarization methods are based on the hierarchical combination of single-document summaries. The automatic evaluation and human study performed show that these methods improve upon current state-of-the-art multi-document summarization systems on two mainstream evaluation datasets, DUC 2007 and TAC 2009. We show a relative improvement in ROUGE-1 scores of 16% for TAC 2009 and of 17% for DUC 2007.

1. Introduction

Many automatic summarization systems have been proposed in order to cope with the growing number of news stories published online. The main goal of these systems is to convey the important ideas in these stories, by eliminating less crucial and redundant pieces of information. In particular, most of the work in summarization has been focused on the news domain, which is strongly tied to events, as each news article generally describes an event or a series of events. However, few attempts have focused on the use of automatic techniques for event classification for summarization systems for the news domain [1]. In fact, most of the work on multi-document summarization are either based on centrality-based [2–5], Maximal Marginal Relevance (MMR) [6–9], and coverage-base methods [1,10–15]. Generally, centrality-based models are used to generate generic summaries, the MMR family generates query-oriented ones, and coverage-based models produce summaries driven by topics or events.

The use of event information in multi-document summarization can be arranged in the following categories: initial **hand-based experiments** [16]; **pattern-based approaches** based on enriched representations of sentences, such as the cases of the work presented by Zhang et al. [15] and by Li et al. [13], which define events using an event key term and a set of related entities, or centrality-based approaches working over an event-driven representation of the input [1], where events are also pattern-based defined; and, **clustering-based** event definition [17].

The major problem of these approaches is that is difficult to relate different descriptions of the same event due to different lexical realizations. In our work, we address this problem by using an event classification-based approach and including event information supported by two different distributed representations of text—the skip-ngram and continuous bag-of-words models [18]. Our event detection and classification framework is based on vector-valued fuzzy sets [19,20]. We evaluate our work using the standard summarization evaluation metric, ROUGE [21]. Moreover, to better understand the impact of using event information, we also perform a human evaluation using the Amazon Mechanical Turk[1].

* Corresponding author at: INESC-ID Lisboa, Rua Alves Redol, 9, 1000-029 Lisboa, Portugal. Tel.: +35 1213100300. Q2
E-mail address: luis.marujo@inesc-id.pt (L. Marujo).

[1] https://www.mturk.com/

http://dx.doi.org/10.1016/j.knosys.2015.11.005
0950-7051/© 2015 Published by Elsevier B.V.

Our main goal in this work was to produce event-based multi-document summaries that are informative and could be useful for humans. The human evaluation shows that our summaries are on average more useful for humans than the reference summaries. While we conducted our experiments in the news domain, our methods are also applicable to other domains, such as opinion and meta-review summarization in consumer reviews [22].

In this document, the next section describes the related work to contextualize the findings obtained in the experimental results. Section 3.2 introduces the Event Detection framework; which is enhanced by the Continuous Skip-gram Model presented in Section 3.3; both are included in a Event-based Multi-Document Summarization framework (Section 3). The experimental results are included and discussed in Section 4. Section 5 details the conclusions and discusses future research directions.

2. Related work

An early attempt at event-based multi-document summarization, proposed by [16], manually annotated events and showed that events are an useful cue for summarization systems. However, manually extracting events is undesirable as if hampers the automation of summarization systems.

Most of the work in automatic summarization concentrates on extractive summarization. In fact, extracting the important content is the first step of a generic summarization system. The extracted information can subsequently be further processed if the goal is to generate *abstracts*. For this case, the important content is generally devised as a set of concepts that are synthesized to form a smaller set and then used to generate a new, concise, and informative text. The alternative goal can also be to generate *extracts* where the identified content consists of sentences that are concatenated to form a summary.

The most popular multi-document summarization baselines follow into one of the following general models: centrality-based [2–4], Maximal Marginal Relevance (MMR) [6–9], and coverage-base methods [1,10–15,23,24].

Traditionally, *centrality-based* models are used to produce generic summaries, the *MMR* family generates query-oriented ones, and *coverage-base* models produce summaries driven by topics or events.

The most popular centrality-based method is the centroid [2] for multi-document summarization distributed in the MEAD framework. Expected n-call@k [7–9] adapted and extended MMR with new similarity and ranking methods.

Concerning the idea of using event information to improve summarization, previous work [1,12–15] defines events as triplets composed by a named entity, a verb or action noun, and another named entity, where the verb/action noun defines a relation between the two named entities. This information is then included in a generic unit selection model, often trying to minimize redundancy while maximizing the score of the important content. Others have tried to use time information and word overload to summarize the same events [25,26]

In our work, we use, not only event information, but also their classification according to ACE [27]; we additionally explore the possibility of using events to filter out unimportant content; and, to our best of our knowledge, we present the first analysis of the impact of using this type of information on multi-document summarization.

Over the past years, the research community has been exploring event detection. The bulk of the event detection work started in the end of 1990s with the Topic Detection and Tracking (TDT) effort [28–31]. The TDT project had two primary tasks: First Story Detection or New Event Detection (NED), and Event Tracking. The objective of the NED task was to discover documents that discuss breaking news articles from a news stream. In the other task, Event Tracking, the focus was on the tracking of articles describing the same event or topic over a period of time. More recent work using the TDT datasets [32–34] on Event Threading tried to organize news articles about armed clashes into a sequence of events, but still assumed that each article described a single event. Passage Threading [33] extends the event threading work by relaxing the one-event-per-news-article assumption. For this purpose, it uses a binary classifier to identify “violent” events in paragraphs.

Even though the TDT project ended in 2004, new event detection research continued. The most well-known example is Automatic Content Extraction (ACE. The goal of ACE research is to detect and recognize events in text. Beyond the identification of events, the ACE 2005 [27] task identifies participants, relations, and attributes of each event. This extraction is an important step towards the overarching goal of building a knowledge base of events [35]. More recent research [36] explores bootstrapping techniques and cross-document techniques augmenting the ACE 2005 with other corpora, including MUC-6 (Message Understanding Conference).

The idea of augmenting the ACE 2005 corpus stems from the low occurrence of some event types in the sentences of the dataset. Most sentences do not contain any event or describe an event that does not exist in the list of event types, which makes the identification of events a complex task. Additional features combined with supervised classifier [37], such as SVM, improved the identification of events. But a more simple and efficient approach based on Fuzzy Logic outperformed the best results. For this reason, we are using it in this work.

As discussed above, events are hard to detect. However, the identification of anomalous events makes the task simpler [38]. Still, determining if two events are the same or are related is, as noted by Hovy et al. [39], an unsolved problem. Even event co-reference evaluation is not a trivial problem [40].

While word embeddings have been used in many NLP tasks [41,42], they have not been used in event detection or summarization to the best of our knowledge. The closest work found is a summarization work that trains a neural network to learn the weights for a small set of features.

Even considering that clustering-based event definition approaches could handle this type of problem, the work of Li et al. [17] models events in a similar way of topics.

3. Event-based multi-document summarization

Our multi-document summarization approach is based on a single-document centrality summarization method, KP-Centrality [43] (Fig. 1). This method is easily adaptable [44] and has been shown to be robust in the presence of noisy input. This is an important feature, since the multiple documents given as input in multi-document summarization are more likely to contain unimportant information compared to single-document summarization.

3.1. From single-document to multi-document summarization

Our goal is to extend the KP-Centrality method for multi-document summarization. The simplest method would be to concatenate all documents and use the single-document method to produce the summary. We shall use this approach as a baseline. This baseline works quite well for a small number of documents, but the performance decreases as the number of documents increases. This means that KP-Centrality has limitations identifying redundant content, such as events, when it is written with different words. Another limitation of the baseline method is to ignore temporal information as more recent news documents tend to contain more relevant information and sometimes include brief references to the past events to provide some context.

To overcome the first limitation, we consider two simple but effective alternative approaches for improving the baseline method. The