Expert Systems with Applications 41 (2014) 6904–6916

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Event graphs for information retrieval and multi-document summarization

Goran Glavaš, Jan Šnajder *

University of Zagreb, Faculty of Electrical Engineering and Computing, Text Analysis and Knowledge Engineering Lab, Unska 3, 10000 Zagreb, Croatia

ARTICLE INFO

Article history:
Available online 19 April 2014

Keywords:
Event extraction
Information extraction
Information retrieval
Multi-document summarization
Natural language processing

ABSTRACT

With the number of documents describing real-world events and event-oriented information needs rapidly growing on a daily basis, the need for efficient retrieval and concise presentation of event-related information is becoming apparent. Nonetheless, the majority of information retrieval and text summarization methods rely on shallow document representations that do not account for the semantics of events. In this article, we present *event graphs*, a novel event-based document representation model that filters and structures the information about events described in text. To construct the event graphs, we combine machine learning and rule-based models to extract sentence-level event mentions and determine the temporal relations between them. Building on event graphs, we present novel models for information retrieval and multi-document summarization. The information retrieval model measures the similarity between queries and documents by computing graph kernels over event graphs. The extractive multi-document summarization model selects sentences based on the relevance of the individual event mentions and the temporal structure of events. Experimental evaluation shows that our retrieval model significantly outperforms well-established retrieval models on event-oriented test collections, while the summarization model outperforms competitive models from shared multi-document summarization tasks.

1. Introduction

The amount of textual data reporting on real-world events (e.g., breaking news, police reports, social media posts) is increasing rapidly on a daily basis. At the same time, there exists the need to obtain information about current and past events, such as finding out more about "*Obama visiting Russia and meeting Putin*" or "*French chateau sale ending in tragedy*". With potential applications ranging from media analysis and tracking to security and intelligence, it has become increasingly important to address such event-oriented information needs. Despite this, many contemporary information retrieval (IR) systems (Castells, Fernandez, & Vallet, 2007; Sarkar, 2012; Turney & Pantel, 2010) still implement or build upon the traditional retrieval models (Ponte & Croft, 1998; Robertson & Jones, 1976; Salton, Wong, & Yang, 1975), which rely on a shallow, bag-of-words representation of documents and keyword-based queries. These models are unable to account for the semantics of events, especially their temporal structure (e.g., "*International aid was sent after the storm ravaged the country*"). Furthermore, considering that numerous textual sources provide information about the same real-world events, the need for aggregating and summarizing the most relevant information has become obvious. Nevertheless, studies on event-based text summarization are rare (Daniel, Radev, & Allison, 2003; Filatova & Hatzivassiloglou, 2004; Li, Wu, Lu, Xu, & Yuan, 2006). This dearth of studies is rather surprising if one considers that news stories primarily describe real-world events (i.e., an event is a dominant information concept in news) (Pan & Kosicki, 1993; Van Dijk, 1985) and that following an event through several newswires is a prototypical application of multi-document summarization (Barzilay, McKeown, & Elhadad, 1999).

While being extensively studied in linguistics for over half a century (Gennari, Sloman, Malt, & Fitch, 2002; Mayo, 1950; Pustejovsky, 1991), it is only in the last decade that events have received significant research attention in information retrieval and natural language processing (NLP) (Allan, 2002; Pustejovsky et al., 2003a). In topic detection and tracking (TDT), a subfield of IR, the goal is to detect documents discussing new events from the real world and to track their development in time. In TDT, an event is vaguely defined as something that happens in a certain place at a certain time (Yang et al., 1999), whereas topics are

* Corresponding author. Tel.: +385 16129871.
E-mail addresses: goran.glavas@fer.hr (G. Glavaš), jan.snajder@fer.hr (J. Šnajder).

http://dx.doi.org/10.1016/j.eswa.2014.04.004
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

considered sets of news stories related by some seminal real world event (Allan, 2002). To identify news stories on the same topic, most TDT approaches rely on traditional vector space models (Salton et al., 1975), as more sophisticated NLP techniques have not yet proven useful for this task. Meanwhile, there have been significant advances in sentence-level event extraction, which focuses on the extraction of linguistic events or *event mentions* evoked by so-called *event anchors*, which are typically predicates (e.g., "*Chinese warship attacked Philippine fishing boats in South China Sea*"). The research on event extraction has built on standardization efforts such as TimeML (Pustejovsky et al., 2003a) and corpora such as TimeBank (Pustejovsky et al., 2003b), and it has been motivated by a number of dedicated shared evaluation tasks (ACE, 2005; UzZaman et al., 2013; Verhagen et al., 2007, Verhagen, Sauri, Caselli, & Pustejovsky, 2010). Despite these recent developments, research in TDT has remained largely isolated from research on event extraction and has thus far failed to profit from sentence-level event processing in NLP.

In this work, we aim to bridge that gap, and we propose event-oriented retrieval and summarization models based on sentence-level event extraction. We argue that the most relevant information in event-oriented texts is the event mentions and the relations in which they stand to each other, while all other information is event-unrelated and may be considered less relevant. Accordingly, we *filter* the event mentions and *structure* them to capture their relationships (at present, we model only temporal relationships). As an example, consider the following event description:

> *Chinese warship attacked Philippine fishing boats in the South China Sea. South China Sea is a home to a myriad of conflicting territorial claims. The attack was provoked by fishermen refusing to leave what Chinese claim to be their territory.*

Only the first and third sentences are relevant to the event, while the second sentence merely provides background information and may be filtered out. Furthermore, the text gives rise to a temporal structure in which events mentioned in the third sentence ("*provoked*", "*refusing*") preceded the event mentioned in the first sentence ("*attacked*").

To adequately capture the semantics of events, we introduce *event graphs*, a novel event-centered document representation based on sentence-level event mentions. In event graphs, vertices denote the individual event mentions extracted from the text, while edges denote the temporal relations between them. We describe an NLP pipeline that combines supervised machine learning and rule-based models for the extraction of event graphs from English text. Building on event graphs, we propose novel models for event-centered information retrieval and multi-document text summarization. The event-centered IR model relies on a semantic comparison between queries and documents by employing graph kernels over event graphs. The event-centered multi-document summarization employs event graphs to assign relevance scores to the individual event mentions and then exploits the structure of event graphs to propagate the relevance to temporally related events.

We demonstrate that our models achieve significant improvements over well-established models on event-centered IR tasks as well as over competing methods for multi-document summarization. The strength of the proposed models stems from the underlying event-oriented information extraction system, which produces graph-based event representations that retain all important aspects of real-world events. The proposed models are therefore particularly suitable for domains that describe real-world events, such as news stories or police reports. On the other hand, the proposed models are not appropriate for domains of descriptive texts (e.g., art reviews) in which event mentions are very rare. The effectiveness of the proposed models is limited by the current state-of-the-art performance of event extraction models. Consequently, even better performance of the proposed retrieval and summarization models is expected with improvements in event-oriented information extraction.

The remainder of the article is organized as follows. In Section 2, we provide an overview of work on event processing in NLP and TDT and its applications in IR and text summarization. Section 3 formalizes an event graph and describes the pipeline for extracting event graphs from text. In Section 4, we present and evaluate the event-centered IR model based on event graphs and graph kernels, while in Section 5 we present and evaluate the event-centered multi-document summarization model. Section 6 concludes the paper and outlines directions for future research.

2. Related research

Following the nature of the work we present in this article, the review of the related research is threefold. We first present the most influential research on event and temporal information detection in NLP and TDT. Second, we give an overview of event-based approaches to information retrieval. Third, we provide an overview of event-based approaches to text summarization.

2.1. Event and temporal information extraction

The introduction of standards for annotating sentence-level events and temporal information (Pustejovsky et al., 2003a) in text and the development of the corresponding datasets (Pustejovsky et al., 2003b) nearly a decade ago marked the beginning of a period of intensive research in event processing in NLP, driven primarily by designated shared tasks (ACE, 2005; UzZaman et al., 2013; Verhagen et al., 2007, 2010). Following the early attempts (Aone & Ramos-Santacruz, 2000; Grishman & Sundheim, 1996; Humphreys et al., 1998), the focus of the Automated Content Extraction (ACE) event extraction tasks were focused on extracting events for specific domains. The tasks included the extraction of event anchors and event arguments as well as event coreference resolution. Ahn (2006) proposed the approaches based on supervised machine learning for all three event-oriented tasks. The first TempEval competition (Verhagen et al., 2007) had three different temporal relation extraction tasks: extraction of relations between events and temporal expressions, between events and document creation time (DCT), and between the main events of adjacent sentences. The second competition (Verhagen et al., 2010) was extended with three additional tasks: the extraction of event anchors, the extraction of temporal expressions, and the recognition of temporal relations between events from the same sentence, where one syntactically dominates the other. The best performance on the anchor extraction task was achieved by the systems based on rather different approaches: Grover, Tobin, Alex, and Byrne (2010) used a rule-based approach in which they filtered head verbs and head nominalizations with WordNet-based attributes, whereas Llorens, Saquete, and Navarro (2010) used supervised machine learning with conditional random fields and a rich set of linguistic features. The best-performing system on the temporal relation extraction task used supervised machine learning with Markov logic networks (UzZaman & Allen, 2010).

Unlike the aforementioned body of work, in which tasks considering events and temporal information were considered in isolation (e.g., temporal relation extraction between pairs of manually labeled event mentions), more recent research has focused on extracting global temporal representation of documents (Bethard, 2013; Bramsen, Deshpande, Lee, & Barzilay, 2006; Kolomiyets, Bethard, & Moens, 2012; UzZaman et al., 2013). Bramsen et al.