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/

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.

(2006) represent documents as temporal directed acyclic graphs (DAGs) of textual segments. They consider only temporal precedence, not temporal overlap or the temporal equivalence of different segments of text. Kolomiyets et al. (2012) represent children's stories as temporal dependency trees. They perform the so-called *temporal parsing* of children's stories using different parsing approaches: a deterministic non-projective shift-reduce parser and a graph-based maximum spanning tree parser. In line with this work, the most recent TempEval evaluation campaign (UzZaman et al., 2013) included a new *temporal awareness* task in which the participating systems had to produce temporal graphs in which vertices denote event mentions and temporal expressions whereas edges denote temporal relations. The most temporally aware system (Bethard, 2013) used a pipeline of machine learning models.

In TDT, the clustering of news stories and the detection of stories describing new events mostly rely on the traditional vector space model (Salton et al., 1975), which represents documents as vectors of words and uses the cosine between the vectors as a measure of document similarity. However, more recent work (Hatzivassiloglou, Gravano, & Maganti, 2000; Kumaran & Allan, 2004; Makkonen, Ahonen-Myka, & Salmenkivi, 2004) acknowledges the importance of specific word classes (e.g., named entities, noun phrases, collocations) and extends the traditional vector space model to consider these separately. Named entities have been recognized as particularly important, as they often identify the participants of an event. Hatzivassiloglou et al. (2000) represent documents using only words that constitute noun phrases or named entities. Makkonen et al. (2004) split the words into four semantic categories – temporal expressions, locations, named entities, and general terms – and construct a separate vector for each of the categories. The recognition of documents that report on the same events can also rely on meta-information associated with news stories, such as DCT, if such information is available. For example, (Atkinson & Van der Goot, 2009) combine DCT with a vector space model, assuming that the documents being far apart in time are less likely to discuss the same events. However, the need for a more structured representation of events has been recognized within the TDT community (Makkonen, 2003). Although traditional TDT approaches predominantly group news into flat clusters, approaches have been proposed for structuring news into hierarchical event threads (Nallapati, Feng, Peng, & Allan, 2004) and tracking the evolution of events (Wei & Chang, 2007; Yang, Shi, & Wei, 2009).

Similar to TempEval-3, in this work we construct a structured event-centered representation of documents. However, the event graphs we propose are semantically richer because, in addition to temporal information (i.e., event anchors and temporal relations), they contain semantic arguments of events (i.e., participants and circumstances). With the event graph providing a structured event-centered document representation, we also address the gap recognized by the TDT community.

2.2. Event-based information retrieval

Although many information needs are event-oriented (and, as such, their semantics can hardly be expressed using keyword-based queries) and there are numerous collections of documents describing real-world events (e.g., breaking news, police reports), there have been very few attempts to exploit events or event-centered structures in information retrieval.

Lin, Yen, Hong, and Cruz-Lara (2007) explicitly focus on event-oriented information retrieval by comparing the semantic roles of the predicate arguments in the query with the semantic roles of the predicates extracted from documents. The semantic roles of the queries need to be annotated manually, whereas all predicate frames from the document are extracted automatically. Query-document matches are recognized based on a comparison of entities occupying the same semantic roles of the matching predicates. Similarly, Kawahara, Shinzato, Shibata, and Kurohashi (2013) compare the predicate-argument structure of the query with the predicate-argument structures extracted from the document, but they extract the query structure automatically. They employ a semantic parser that uses dependencies from PropBank (Kingsbury & Palmer, 2002) and they demonstrate that ranking based on semantic roles outperforms ranking based on syntactic dependencies.

Although they account for syntactic variation by relying on semantic representations, neither of the above-described approaches accounts for queries consisting of multiple predicates, such as those expressing temporal relations between events (e.g., "*elections before protests*"). In this work, we present an event-oriented IR model that is capable of exploiting the structure between events (i.e., temporal order) in both the documents and queries and can thus account for queries with underlying temporal semantics involving multiple events.

Taking a broader view, many IR models have been proposed that use structured document representations that are not based on events, e.g. (Pai, Chen, Chu, & Chen, 2013; Sbattella & Tedesco, 2013). Pai et al. (2013) structure documents as sets of subject-predicate-object triples, the so-called *content maps*. The similarity between two documents is determined by comparing their content maps, while the triples themselves are compared by comparing the words from the corresponding syntactic categories (e.g., a subject from one triple is compared to a subject in another triple). Although in some aspects similar to our work, triples that Pai et al. consider in many cases do not correspond to event mentions. Furthermore, their model only accounts for the similarity between pairs of documents and not between a document and a query. As regards the performance, no firm conclusions can be drawn because of the absence of comparison with traditional IR paradigms. Sbattella and Tedesco (2013) represent documents at two levels: the conceptual and lexical level. At the conceptual level, they tag the documents with the concepts from a domain ontology, whereas at the lexical level they associate the words with synonym sets from WordNet. Their model is capable of handling keyword-based queries but also more complex natural language queries. However, at the conceptual level their model relies on a domain ontology, which is often not available in ad hoc information retrieval settings. Unlike the event-based IR model we present, the model of Sbattella and Tedesco (2013) employs a shallow-structured document representation that does not incorporate any event-specific semantics.

2.3. Event-based text summarization

Despite the fact that (multi-) document summarization approaches predominantly focus on summarizing newswire texts and that events are the primary concept of news (as news describes real-world events), very few attempts have been made towards realizing event-oriented text summarization.

Event-based multi-document summarization was first proposed by Daniel et al. (2003) who, following the TDT definition of an event (Allan, 2002), selected sentences based on relevance for one or more sub-events of the topic at hand. Human judges manually determined the sub-events of a topic and assigned to each sentence a relevance score for each sub-event. They show that the algorithm that selects sentences with the highest sum of scores over all sub-events produces the most informative summaries. However, this approach is not fully automated, as it requires significant human effort for annotating the relevance of sentences across all topic sub-events. It also only considers document-level events,

neglecting the information originating from sentence-level events (i.e., event mentions).

Filatova and Hatzivassiloglou (2004) shift the focus to sentence-level events by evaluating sentences according to co-occurrence statistics between named entities and verbs (or action nouns). They adopt a definition of an event mention as any co-occurrence of salient named entities with a verb or an action noun in between. However, this is a rather narrow and imprecise account of an event mention, as informationally important event mentions need not necessarily involve named entities (e.g., "*The rebels then detonated the bomb, wounding at least dozen people*").

Li et al. (2006) extend the work of Filatova and Hatzivassiloglou (2004) by building the co-occurrence graph between named entities and event terms. They assign initial relevance scores to each of the named entities and event terms and then run the PageRank algorithm (Page, Brin, Motwani, & Winograd, 1998) to determine the context-dependent relevance of named entities and event terms. Finally, they compute the relevance of a sentence by summing up the relevances of the named entities and event terms it contains. Their work lacks an automated extraction of event mentions and relations between events other than co-occurrences.

There are a number of approaches to multi-document summarization that first semantically annotate the sentences and then assign relevance scores to sentences based on these annotations, e.g. (Atkinson & Munoz, 2013; Baralis, Cagliero, Jabeen, Fiori, & Shah, 2013). Atkinson and Munoz (2013) begin by assigning rhetorical roles to sentences, and then give preference to sentences with certain roles (e.g., Antecedent) over sentences with other roles (e.g., Conclusion). However, the relevance scores are computed in a traditional manner, by summing the TF-IDF weights assigned to the individual words. Baralis et al. (2013) semantically annotate sentences with concepts from the popular ontological knowledge base Yago and then score the sentences based on the concepts they contain. Next, they iteratively select the subset of the most informative sentences by employing a variant of the maximal marginal relevance strategy. Conceptually, their approach is similar to the event-based summarization model we propose in this article. However, instead of recognizing concepts from a knowledge base, we extract event mentions and assign scores to sentences based on informativeness of events they contain.

A related strand of research is concerned with update summarization (Dang & Owczarzak, 2008; Du, Guo, Zhang, & Cheng, 2010; He, Qin, & Liu, 2012; Yan et al., 2011): focused summarization of novel information from an on-line stream of documents about the same topic (i.e., documents describing the same seminal event) that evolves over time. Update summarization assumes that the reader is already familiar with the previous documents from the same topic. To the best of our knowledge, update summarization systems do not explicitly leverage event-oriented knowledge by considering sentence-level events.

Canhasi and Kononenko (2014) propose a query-focused graph-based multi-document summarization in which, in addition to links denoting the similarity between sentences, they introduce links between sentences and the query denoting how similar individual sentences are to the query. The importance of the sentences is then determined by a matrix factorization method called *weighted archetypal analysis*. Our approach is similar to that of Canhasi and Kononenko in that we also use a graph-based structure (event graphs) to refine the importance scores of sentences.

In this work, we present an extractive multi-document summarization model based on the extraction of sentence-level event mentions and the temporal structure of documents. Our model is non-focused, i.e., the summarization is not driven by a specific query or a timeline constraint. More precisely, our model estimates the relevance scores of individual events using temporal relations between them (i.e., an event is more relevant if it occurs directly before or after a very relevant event) instead of plain co-occurrences or similarities between sentence vectors.

3. Construction of event graphs

We structure news stories as event graphs in which nodes denote event mentions consisting of anchors and arguments, whereas edges denote temporal relations between event mentions. The construction is arranged in a three-stage pipeline that combines machine learning and rule-based extraction methods. More precisely, we use the pipeline to (1) extract event anchors using a supervised model, (2) extract event arguments using a set of hand-crafted rules, and (3) extract and classify temporal relations between pairs of events with a supervised model. We begin with a formal definition of an event graph and then describe the three extraction stages.

3.1. Event graph

We define an event graph G as a tuple $G = (V, E, A, m, r)$, where V is the set of vertices, E is the set of undirected edges, A is the set of directed edges (arcs), $m : V \rightarrow M$ is a vertex-labeling function mapping the vertices to event mentions, and $r : E \rightarrow R$ is the edge-labeling function, assigning temporal relations to edges. Undirected edges model the symmetric temporal relations (Overlap and Equal), while directed edges model the asymmetric temporal relations (Before and After).

In this work, we extract arguments of four coarse-grained types (Agent, Target, Time, and Location). The motivation for this is twofold. First, we consider these types to be informationally the most relevant for any real-world event. Second, by restricting it to a small number of generic argument types, we make the extraction more robust, avoiding the performance issues typically associated with fine-grained semantic role labeling approaches. Similarly, aiming for a robust recognition of temporal relations, we work with four coarse-grained relation types: Before, After, Equal, and Overlap. The first three types are adopted directly from Allen (1983). The Overlap type covers a number of Allen's relations – Overlaps, Starts, During, and Finishes – as these are generally too fine-grained to be reliably detected in text (Verhagen et al., 2010).

As an example, consider the following news story snippet (event anchors are shown underlined):

Egyptian forces have clashed with militants after entering a town near Cairo. Soldiers went into Kerdasah at about 05:30 local time and targeted terrorist hotbeds. Meanwhile, militants shot dead Gen. Nabil Farag, state media said.

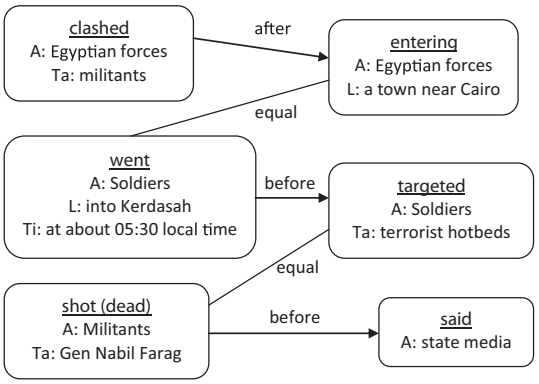


Fig. 1. Example event graph. Event mentions are shown in rounded rectangles (with anchors underlined and arguments labeled with argument types). Temporal relations are shown as arrows (Before and After) and lines (Equals and Overlap).

Table 1
Representative argument extraction patterns.

Name	Example	Dependencies	Argument type
Nominal subject	"***Military*** *confronted protesters*"	*nsubj (confronted, Military)*	Agent
Direct object	"*The government rejected* ***negotiations*** *with protesters*"	dobj (rejected, negotiations)	Target
Prepositional object	"*Protesters demonstrated again* ***on Sunday***"	*prep (protested, on)* and *pobj (on, Sunday)*	Time
	"*The protest* ***in Cairo***"	*prep (protest, in)* and *pobj (in, Cairo)*	Location
	"*The protest* ***against Muslim Brotherhood***"	*prep (protest, against) and pobj (against, Brotherhood)*	Target
Noun compound	"***Cairo*** *protests*"	*nn (protests, Cairo)*	Location
	"***Tuesday*** *demonstrations*"	nn (demonstrations, Tuesday)	Time
	"***EU*** *sanctions*"	*nn (sanctions, EU)*	Agent

The corresponding event graph is shown in Fig. 1.

3.2. Anchor extraction

An event anchor is a word that best captures the core meaning of an event (Ahn, 2006). Anchor extraction is performed by identifying the tokens in text that correspond to event anchors (e.g., "*shot*" and "*said*" in "*Militants shot Gen Nabil Farag, state media said*"). Because our focus is on informationally relevant events, we chose to extract only anchors of *factual* event mentions (mentions of events that indeed happened in the real world), thereby disregarding the negated, hypothetical, and uncertain event mentions. We frame anchor extraction as a classification task and use a discriminative model (a logistic regression classifier) to identify the event anchors. The model uses the following sets of features:

- *Lexical and part-of-speech features* – word, lemma, stem, and part-of-speech tag of the token and its surrounding tokens (two tokens to the left and right);
- *Syntactic features* – the set of dependency relations of the token, its chunk type (e.g., verb phrase, noun phrase), and features denoting whether the token is the head of a nominal subject or a direct object. We computed three features based on the output of the Stanford dependency parser (De Marneffe, MacCartney, & Manning, 2006);
- *Modifier features* – modal modifiers (e.g., *might*), auxiliary verbs (e.g., *been*) and negations of the token. These features are particularly useful for discriminating factual from non-factual event mentions.

To train and test the model, we used the EVENTA corpus[1] consisting of 759 news articles (330 K tokens), manually annotated for factual event mentions. We use 70% of the documents for training and 30% for testing. The anchor recognition model achieves a precision of 83%, a recall of 77%, and an F1-score of 80%. The performance of our anchor extraction model on the standard TimeBank dataset (Pustejovsky et al., 2003b) is 81% F1-score, which is comparable to state-of-the-art approaches for event anchor extraction (Grover et al., 2010; Llorens et al., 2010).

3.3. Argument extraction

Our argument extraction approach relies on a rich set of unlexicalized, dependency-based syntactic patterns introduced by Glavaš and Šnajder (2013a). The set consists of 13 extraction patterns, the most representative of which are shown in Table 1. Some extraction patterns serve only to identify an argument, while additional processing is required to resolve its semantic type (for instance, a prepositional object can be an argument of the TIME, LOCATION, or TARGET type). To this end, we employ a named entity recognition (e.g., if an argument is a named entity of type LOCATION, then the argument is declared to be locative), temporal expression recognition (if an argument is a part of a temporal expression, it is considered to be temporal), and a measure of WordNet-based semantic similarity with temporal and locative concepts (e.g., *location*, *geographical_area*, and *facility* for locations, and, e.g., *time*, *duration*, *time_period* for temporal arguments). For named entity recognition, we use the system of Finkel, Grenager, and Manning (2005); for temporal recognition, we use the system of Chang and Manning (2012); and for WordNet-based semantic similarity, we use the measure proposed by Wu and Palmer (1994).

We tested our rule-based argument extraction approach on the test portion of the EVENTA corpus. The per class F1-score of the model is as follows: AGENT – 88.0%, TARGET – 83.1%, TIME – 82.3%, and LOCATION – 67.5%. The extraction performance for all types except LOCATION is satisfactory (and, though not directly comparable, is above the 70% F1-score typical for semantic role labeling systems). The performance for LOCATION is lower than the performance for TIME because LOCATION arguments (e.g., "*in the car*", "*behind the hospital*", "*north of San Francisco*") exhibit a much larger lexical variation than TIME arguments (mostly names of week days, months, and other easily recognizable expressions, such as "*today*" and "*three years ago*").

3.4. Temporal relation extraction

Temporal relation extraction is performed in two steps. We first use a classifier to identify pairs of event mentions between which a temporal relation can be established. We then classify the temporal relation between event mentions m_1 and m_2 (where m_1 precedes m_2 in text) into one of four types: BEFORE, AFTER, OVERLAP, and EQUAL. For both tasks, we use logistic regression models with the following sets of features:

- *Position features* – the set of features that measure the distance between event anchors (in number of tokens) and their relative position (same sentence, adjacent sentences, adjacent event mentions);
- *Lexical features* – word, lemma, stem, and POS of both event anchors as well as a feature indicating whether the word forms of anchors are the same, a feature indicating the semantic similarity of the anchors (Wu & Palmer, 1994), and the bag-of-words (BoW) between the anchors;
- *Syntactic features* – the set of features based on the dependency parses of sentences. Syntactic features are computed only for a pair of event mentions from the same sentence. They include the syntactic path between the anchors (dependency relation labels on the syntactic path between the anchors), features indicating whether one anchor syntactically dominates the other, and features indicating whether one of the events is a predicate of an adverbial clause governed by the other event;

[1] Obtainable from http://takelab.fer.hr/data/grapheve/.

- *Modifier features* – the set of features that includes all features that describe the modal, auxiliary, negation, and determination modifiers of both event anchors.

To train and test our models, we use the TimeBank corpus (Pustejovsky et al., 2003b), with 70% of the documents used for training and 30% for testing. The models achieve a performance comparable to the state of the art (Verhagen et al., 2010): relation identification achieves an F1-score of 79% (with precision and recall balanced at 79%), while relation classification achieves an overall F1-score of 57%, macro-averaged over all classes (BEFORE – 66%, AFTER – 59%, OVERLAP – 35%, and EQUAL – 68%). This performance is also comparable to that of state-of-the-art models for temporal relation extraction (Llorens et al., 2010; UzZaman & Allen, 2010).

4. Event-centered information retrieval

The main idea behind our event-centered information retrieval model is to represent both the documents and the query as event graphs. This effectively filters out all event-unrelated content, leaving event mentions as the only pieces of information relevant for event-oriented information needs. Furthermore, event graphs introduce a temporal structure between the individual event mentions, which mirror the structure of topic-level events. To measure the similarity between a query and a document, we compare their corresponding event graphs using graph kernels, and we rank the documents according to the obtained similarity scores. By using graph kernels, we are able to account for both the similarity of event mentions as well as the similarity of temporal structures.

We continue with a description of the event graph-based retrieval model, beginning with event graph similarity based on graph kernels. We then describe the experimental evaluation of the model.

4.1. Event graph kernels

Graph kernels (Borgwardt, 2007) are an expressive and efficient (polynomial time complexity) alternative to traditional graph comparison techniques (e.g., subgraph isomorphism algorithms). Despite the obvious advantages, graph kernels have so far been utilized in only a few NLP and IR tasks, such as question answering (Suzuki, Isozaki, & Maeda, 2004; Suzuki, 2005) and cross-lingual IR (Noh, Park, Yoon, Lee, & Park, 2009). In our previous work (Glavaš & Šnajder, 2013b), we demonstrated that graph kernels over event graphs can be used successfully for measuring event-centered similarity among event-oriented documents for the purpose of identifying topically related news stories (i.e., news stories discussing the same events).

In this work, we employ two types of graph kernels: a *product graph kernel* (Gärtner, Flach, & Wrobel, 2003) and a *weighted decomposition kernel* (Menchetti, Costa, & Frasconi, 2005). We chose these particular kernels because their general forms have intuitive interpretations for event matching and can be easily adjusted to fit our needs for event-centered text comparison.

The computation of event graph kernels relies on the ability to identify nodes in two graphs that represent the same real-world events – a task that amounts to resolving cross-document event coreference. To this end, we use an event coreference resolution model based on the comparison of event anchors and coarse-grained arguments, as we proposed in our previous work (Glavaš & Šnajder, 2013a). This model achieves an F-score of 67% (79% precision and 57% recall) on the cross-document section of the EventCorefBank dataset (Bejan & Harabagiu, 2008). In what follows, we use $coref(m_1, m_2)$ to represent the output of the event coreference model, where the value is 1 if the model predicts that mentions m_1 and m_2 co-refer and 0 otherwise.

4.1.1. Product graph kernel

A product graph kernel (PGK) counts the common walks between the two input graphs (Gärtner et al., 2003). The kernel score is computed on the graph product of the input graphs. The product of two labeled graphs, G and G', denoted $G_P = G \times G'$, is a graph that has the following vertex set

$$V_P = \{(v, v') | v \in V_G, v' \in V_{G'}, \delta(v, v')\}$$

where $\delta(v, v')$ is a predicate that holds if and only if vertices v and v' have the same labels (Hammack, Imrich, & Klavžar, 2011). Given that our input graphs $G = (V, E, A, m, r)$ and $G' = (V', E', A', m', r')$ are event graphs whose vertices denote event mentions, we use event coreference as the label-matching predicate, i.e., $\delta(v, v') = coref(m(v), m'(v'))$. The edge set of the graph product is determined by the product type. In this work, we experiment with two product types: *tensor product* and *conormal product*. An edge of the tensor product is created only for pairs of matching edges between input graphs G and G'. Two edges match if they denote the same temporal relation between pairs of matching vertices (i.e., vertices denoting coreferring event mentions):

$$E_P = \{((v, v'), (w, w')) \in V_P \times V_P | (v, w) \in E_G, (v', w') \in E_{G'}, \\ r(v, w) = r'(v', w')\}$$

An edge of the conormal product is introduced for every pair of matching vertices between which there is an edge in at least one of the input graphs. If there is a corresponding edge in both graphs, the requirement of matching edge labels (i.e., the same temporal relation) is not imposed:

$$E_P = \{((v, v'), (w, w')) \in V_P \times V_P | (v, w) \in E_G \vee (v', w') \in E_{G'}\}$$

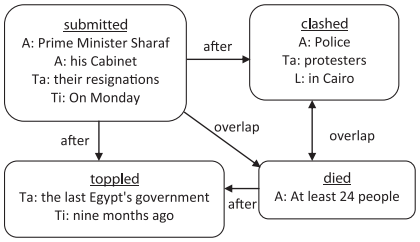
Thus, a conormal product may compensate for temporal relations erroneously omitted in the extraction of the input graphs. Put differently, if we identify two pairs of coreferent event mentions, $(m(v), m'(v'))$ and $(m(w), m'(w'))$, but identify a temporal relation only between mentions in one of the graphs, say $m(v)$ and $m(w)$, we then assume that the same temporal relation holds between their coreferent counterparts in the other graph, $m'(v')$ and $m'(w')$. However, in cases where this assumption is incorrect (i.e., the edge that exists in one of the graphs is spurious), the conormal product graph will contain spurious edges that do not mirror the true overlap between the input graphs.

As an example, consider the following two news snippets, whose corresponding event graphs, tensor and conormal products are shown in Fig. 2:

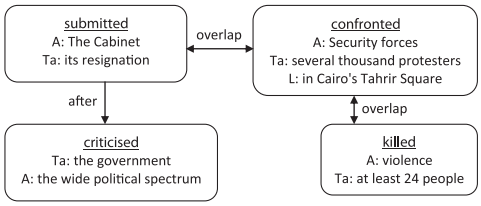
Story 1: *Prime Minister Sharaf and his Cabinet have submitted their resignations to the ruling military council on Monday after police clashed protesters in Cairo third day in a row. At least 24 people have died since the last Egypt's government was toppled nine months ago.*

Story 2: *The Cabinet has submitted its resignation to the ruling military council after the government has been consistently criticized by the wide the political spectrum. Security forces confronted Monday several thousand protesters in Cairo's Tahrir Square in the third straight day of violence that has killed at least 24 people.*

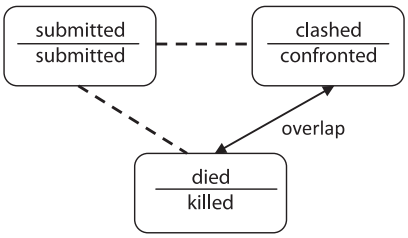
Once the product graph is constructed, the PGK score can be computed efficiently. Let A_P be the adjacency matrix of the product G_P built from input graphs G and G'. The kernel score is then computed as follows:



(a) Event graph G (Story 1)



(b) Event graph G' (Story 2)



(c) Product graph G_P

Fig. 2. Example event graphs and their products (tensor and conormal). The conormal product has three edges (solid and dashed lines), while the tensor product has only one edge (solid line).

$$k_P(G,G') = \sum_{i,j=1}^{|V_P|}\left[\sum_{k=0}^{\infty}\lambda^k A_P^k\right]_{ij} = \sum_{i,j=1}^{|V_P|}[(I-\lambda A_P)^{-1}]_{ij}$$

The kernel score (the sum) converges when $1/t$, where t is the maximum degree of a vertex in the product G_P (Borgwardt, 2007). In all our experiments, we set λ to the standard value of $1/(t+1)$.

4.1.2. Weighted decomposition kernel

In contrast to PGKs, which essentially count common walks between two graphs, a weighted decomposition kernel (WDK) compares arbitrary subgraphs between input graphs according to a matching predicate (Menchetti et al., 2005). Each pair of subgraphs, in this context referred to as *selectors*, contributes to the overall kernel score according to the similarity of the contexts in which the selectors occur.

Let $S(G)$ be the set of all pairs $(s,\mathbf{z})$, where s is the selector (subgraph of interest) and $\mathbf{z}$ is the vector of contexts of s. The general form of the kernel for two input graphs G and G' (Menchetti et al., 2005) is as follows:

$$k_{WD}(G,G') = \sum_{\substack{(s,\mathbf{z})\in S(G)\\(s',\mathbf{z}')\in S(G')}} \delta(s,s')\sum_{d=1}^{D}\kappa_d(z_d,z'_d)$$

where δ is a selector-matching predicate and κ_d is a kernel that measures the similarity of the contexts of matching selectors. For our event graphs, we define selectors to be the individual vertices. The rationale for this is twofold:

1. For individual vertices, we can use the coreference of event mentions as our selector-matching predicate out of the box (i.e., we define the predicate $\delta(v,v')$ for two vertices $v\in G$ and $v'\in G'$ to hold if and only if the corresponding event mentions $m(v)$ and $m'(v')$ co-refer);
2. Choosing subgraphs larger than individual vertices as selectors would demand a more complex and less intuitive selector-matching predicate δ.

For each selector vertex v we define one context Z_v (i.e., the vector of contexts $\mathbf{z}$ has only one component) as a subgraph of G that contains the immediate neighborhood of v (i.e., v and all its immediate neighbors). In other words, given an event mention, we consider as its context all event mentions that are in a direct temporal relation with it. Thus, we compute the WDK between event graphs G and G' as follows:

$$k_{WD}(G,G') = \sum_{v\in V_G, v'\in V_{G'}} coref(m(v),m'(v'))\kappa(Z_v,Z'_{v'})$$

where $\kappa(Z_v,Z'_{v'})$ is the *context kernel* measuring the similarity between the context Z_v of selector $v\in G$ and the context $Z'_{v'}$ of selector $v'\in G'$. The context kernel κ counts the number of coreferent mention pairs found between the contexts, normalized by the size of the larger context:

$$\kappa(Z_v,Z'_{v'}) = \frac{\sum_{w\in V_{Z_v}, w'\in V_{Z'_{v'}}} coref(m(w),m'(w'))}{\max(|V_{Z_v}|,|V_{Z'_{v'}}|)}$$

The rationale behind such a design of the context kernel is that a pair of coreferent event mentions $m(v)$ and $m'(v')$ should contribute to the overall kernel score proportionally to the number of pairs of coreferent mentions $m(w)$ and $m'(w')$ that are in temporal relation with v and v', respectively.

4.2. Experimental setup

The standard IR evaluation paradigm relies on a test collection consisting of documents, queries, and relevance judgments. To the best of our knowledge, no standard test collection is available for event-based information retrieval. Thus, to evaluate our graph-based event-centered models, we decided to build collections on our own. We created two test collections, each containing 50 queries. The first is a mixed topic collection (MixedTopic) of 25,948 news stories covering various topics. The second is a topic-specific collection (OneTopic) containing 1387 news stories related to the Syria crisis. Documents for both collections were obtained through the news clustering service EMM News Brief,[2] which crawls news stories from numerous web sources and clusters them topically.

We asked a human annotator to derive 50 queries for each of the collections as follows: (1) randomly select a document from the collection, (2) read the document thoroughly and with understanding, and (3) compile a query consisting of at least two event mentions such that the selected document is relevant for this query. In effect, the annotator was asked to create a query representing a very short abstractive summary of a randomly selected document (or a salient part thereof). Example queries and the document snippets from which they originate are given in Table 2.

Because annotating the relevance of each document for each query would be unfeasible, we relied on a commonly used pooling method to reduce the number of relevance judgments. We used two baseline IR models, a TF-IDF weighted vector space model (VSM) and a unigram language model (LM) to pool the documents for each query. Our graph-based retrieval models were not used for

[2] http://emm.newsbrief.eu.

Table 2
Example queries with respective news snippets.

Collection	Query	News story snippet
MixedTopic	*United Russia won the elections triggering large demonstrations*	*The pro-Kremlin United Russia party won less than 50 per cent of the vote, a steep fall from its earlier majority, according to preliminary results. But opposition parties and international observers said the vote was marred by widespread reports of vote-rigging. Thousands of security forces were out in the Russian capital and helicopters roamed the sky Wednesday, a show of force following protests over scandal-marred elections that saw Prime Minister Vladimir Putin's party struggle to keep a majority.*
	The bank lost billions in trading and its stocks fell after the loss has been announced.	*A surprise $2 billion trading loss by a division of JPMorgan Chase triggered calls for tougher regulation of banks. Stock in the bank, the largest in the United States, lost 8 percent of its value on Wall Street immediately after the loss announcement, and other American and British banks suffered heavy losses as well.*
OneTopic	*Dozens of bodies were ditched in the streets after vindicatory attacks against Sunnis*	*Dozens of bodies were dumped in the streets of a Syrian city at the heart of the country's nearly 9-month-old uprising, a grim sign that sectarian bloodshed is escalating as the country descends further toward civil war. Up to 50 people were killed in Homs on Monday, but details about what happened in Syria's third-largest city only came to light Tuesday with reports of retaliatory attacks pitting members of the Alawite sect against Sunnis.*
	The minister met with the Syrian opposition and advocated them to unite.	*British Foreign Secretary William Hague on Monday called on Syrian opposition groups to "unite" against Syrian President Bashar al-Assad. Hague made the statement after meeting with Syrian opposition representatives in London on Monday.*

pooling due to the large preprocessing effort. Note that excluding graph-based models from contributing to the pool only favors baseline models because the pool-based IR evaluations are biased towards models contributing to the pool (Büttcher, Clarke, Yeung, & Soboroff, 2007). We decided to pool 75 documents with each of the baseline models because (1) most EMM clusters contain less than 50 documents (which gives, at most, 50 relevant documents for most queries) and (2) we wanted to obtain a very good recall estimate even for queries originating from rare clusters containing over 50 documents. The average pool size per query is 95 news stories.

A single annotator judged the relevance of pooled documents for each query. Initially, we asked a second annotator to annotate the relevance for two randomly chosen queries on which we observed perfect agreement between the annotators. This confirmed our intuition that determining relevance for event-centered queries is not a difficult task (unlike determining relevance for keyword-based queries), and it allowed us to proceed with a single relevance judgment per document. The average number of documents relevant for a query is 12 and 8 in the MixedTopic and OneTopic collections, respectively.

4.3. Results and discussion

To put our results into perspective, we compare our graph-based retrieval models to baseline modes from three traditional retrieval paradigms: (1) vector space retrieval models – TF-IDF-weighted cosine VSM; (2) language models – Hiemstra's (2001) language model; and (3) probabilistic models – the two best-performing models from the Divergence from Randomness framework (In_expC2 and DRF_BM25) (Amati, 2003). We used the implementation of these models within the Terrier IR platform.[3]

The performance of baseline and graph-based models, evaluated in terms of the standard IR evaluation metric, namely, the mean average precision (MAP), is given in Table 3. All our graph-based models significantly outperform all considered baselines ($p < 0.01$ for the tensor PGK model and $p < 0.05$ for the conormal PGK and WDK models; significance tested using paired student's t-test), thus supporting the initial hypothesis that structured event-centered document representation is beneficial for retrieving documents relevant for event-oriented information needs. Although the model employing the tensor product (Tensor PGK) performs better than the models using the conormal product (Conormal PGK) and weighted decomposition (WDK), these differences are not statistically significant ($p < 0.05$). Nonetheless, the fact that the Conormal PGK does not outperform the Tensor PGK implies that the conormal product introduces spurious edges into the product graph more often than it remedies for incorrect extractions of temporal relations between events. We could compensate for this by trading off recall for precision in the temporal relation extraction model.

The performance on the OneTopic collection is consistently lower than on the MixedTopic collection across all models. This is expected because the documents from the OneTopic collection are more similar to each other, as they contain stories about the same topic, making it more difficult to distinguish the relevant from the non-relevant documents. However, the decrease in performance on MixedTopic collection is much smaller for the graph-based models (17% for Conormal PGK, 19% for Tensor PGK, and 20% for WDK) than for the baseline models (e.g., 42% for DFR_BM25). This indicates that graph-based event-centered models are particularly suitable for event-based retrieval over topically focused collections.

The advantage of event graphs over traditional IR models is that they filter only the event-related information and temporally structure this information. It is therefore interesting to analyze how much each of these two aspects contributes to the overall model performance; in particular, we are interested in whether event-oriented information filtering alone suffices to improve the retrieval performance. To this end, we experiment with a model that extracts only the event-oriented information but does not model the temporal relations between them (NoStruct model). The NoStruct model counts the number of pairs of coreferent event mentions (this corresponds to number of vertices in the product graphs) between the query and the document, normalized by the number of event mentions in the document. The performance of the NoStruct model is shown in the last line of Table 3. The model is outperformed by all kernel-based models (difference is statistically significant for tensor PGK at $p < 0.05$), yet performs better than the baseline models (difference is significant for OneTopic collection at $p < 0.05$). This demonstrates that event graph models owe their success to both event-centered filtering and temporal structuring: using filtering alone already outperforms the baselines, but a substantial performance gain is achieved only when combining filtering and temporal structure.

[3] http://terrier.org.

Table 3
Document retrieval performance (MAP). (Best results in each group are shown in bold.)

		Test collection	
	Model	MixedTopic	OneTopic
Baselines	TF-IDF VSM	0.335	**0.199**
	Hiemstra LM	0.300	0.175
	In_expC2	**0.341**	0.188
	DFR_BM25	0.332	0.192
Event graphs	Tensor PGK	**0.502**	**0.407**
	Conormal PGK	0.434	0.359
	WDK	0.449	0.358
	NoStruct	0.374	0.303

5. Event-centered multi-document summarization

Event-oriented text abounds with information about events (e.g., the individual event mentions, event participants, locations) and the relations among them. On the other hand, event-oriented texts typically also contain a non-negligible amount of supporting text, most often in the form of elaborations and background facts (e.g., "*South China Sea is a home to a myriad of conflicting territorial claims*"). Background descriptions do not relate to any of the concrete real-world events and thus do not constitute the gist of the narrative.

Building on these insights, we present a multi-document summarization model that filters the text based on the relevance of event mentions. We rely on a reasonable assumption that the relevance of a sentence correlates with the relevance of the event mentions it contains and that the latter can be estimated based on the information provided by an event graph. The proposed model is *extractive* (summaries are built by extracting sentences from the original texts) and *non-focused* (the summarization is not guided by a specific information need). We proceed with a description of the model, followed by a description of the experimental setup and the results.

5.1. Event-centered summarization model

The summarization model is shown in Algorithm 1. The initial step is the construction of event graphs for all documents from a group of topically related documents (line 3 of the algorithm). We then compute the relevance score of each event mention from the event graph based on three criteria: (1) the importance of the event's participants, (2) the informativeness of the event, and (3) the temporal relations among the events. This is performed in two steps: we first compute the participants' importance score (line 10) and the event informativeness score (line 11). We then use the temporal structure of the event graphs to refine both scores using the PageRank algorithm (lines 13 and 14). We rank the event mentions according to the refined scores and compute the final score for each event mention as the sum of its ranks in both rankings (lines 18 and 19). After obtaining the final scores for the individual event mentions, we compute the scores for each sentence by summing the scores of the event mentions it contains (lines 20–25). To avoid redundancy, we additionally cluster the selected sentences based on semantic similarity (line 26). Finally, we select the representative sentences from each semantic cluster to constitute the summary (lines 27–32). We next describe the individual steps in more detail.

Algorithm 1. Summarize (D, l)

1: **Input**: Document group D; summary length l
2: **Initialize**:
3: G_d – event graph for document d in document group D
4: $E(D)$ – set of all event mentions in document group D
5: $E(d)$ – set of all event mentions in document d
6: $Sent(D)$ – set of all sentences in document group D
7: **Summarize**:
8: $P_{scores} \leftarrow \emptyset; I_{scores} \leftarrow \emptyset$;
9: **for each** event e in $E(D)$ **do**
10: $P_{scores}[e] \leftarrow S_P(e)$
11: $I_{scores}[e] \leftarrow S_I(e)$
12: **for each** document $d \in D$ **do**
13: $P_{scores} \leftarrow$PageRank $(G_d, E(d), P_{scores})$
14: $I_{scores} \leftarrow$PageRank $(G_d, E(d), I_{scores})$
15: $R_P :=$sort $(E(D), P_{scores})$
16: $R_I :=$sort $(E(D), I_{scores})$
17: $S_{scores} \leftarrow \emptyset$
18: **for each** event e in $E(D)$ **do**
19: $S_{scores}[e] \leftarrow$ rank (e, R_P) + rank (e, R_I)
20: $Sent_{scores} \leftarrow \emptyset$
21: **for each** document d in D **do**
21: **for each** sentence s in d **do**
22: $Sent_{scores}[s] \leftarrow 0$
23: **for each** event e in s **do**
24: $Sent_{scores}[s] = Sent_{scores}[s] + S_{scores}[e]$
25: $C \leftarrow$cluster $(Sent(D))$
26: $CR \leftarrow$representatives $(C, Sent_{scores})$
27: $CR \leftarrow$sort_reversed $(CR, Sent_{scores})$
28: $summary \leftarrow \emptyset$
29: **do**
30: $summary \leftarrow summary \cup$ take_first (CR)
31: **while** length $(summary) \leqslant l$
32: **Output**: *summary*

5.1.1. Importance of event participants

The first score we compute for each event mention is the importance of its participants given the topic (we refer to this score as S_P). We consider participants to be named entities that occur as event arguments of the AGENT or TARGET type. Intuitively, participants that occur more frequently are more relevant for the topic. On the other hand, participants that occur frequently in only a small number of documents are likely to be less relevant under the traditional summarization assumption that information present in all (or most) documents is the most relevant. Based on these observations, we estimate the relevance of a named entity by taking into account its overall frequency in the group of documents as well as the number of documents in the group that contain this named entity.

Let D be the group of topically related documents we want to summarize, d an individual document, and e an individual event mention. Given event e, we compute the importance of its participants as follows:

$$S_P(e) = \sum_{p \in P(e)} \sum_{d \in D} \#(p,d) \cdot \frac{|\{d \in D | \#(p,d) > 0\}|}{|D|}$$

where $P(e)$ is the set of all participants of e, p is the individual participant ($p \in P(e)$), and $\#(p,d)$ is the frequency of participant p in document d.

It should be mentioned that incorporating the importance of named entities directly into a measure of informational salience of sentences has been proposed in the literature (e.g., Ge, Huang, & Wu, 2003; Ouyang, Li, Li, & Lu, 2011; Saggion & Gaizauskas, 2004). However, the key distinguishing feature of our model is that we only consider the named entities participating in events. By linking named entities to events, we are able to avoid extracting event-unrelated background sentences containing otherwise relevant named entities.

5.1.2. Event informativeness

Although the participants of events are rather indicative of their relevance, some event mentions can still be very relevant for the topic even if they have no named entities as arguments (e.g., "*Insurgents launched a massive attack early in the morning*"). Conversely, some event mentions containing frequently mentioned participants may not be very relevant for the topic (e.g., "*Obama and Putin took the photograph together at the Lough Erne Resort*", with respect to the topic of a political summit).

To address this point, we compute for each event mention a score that aims to capture the general informativeness of an event mention regardless of its participants. We do this by comparing the informativeness of an event mention's constituent words within the group of topically related documents against their informativeness within a large general-topic corpus. We use the difference of relative frequencies of a word in these two document sets as a measure of how relevant each word is with respect to the topic. There are two assumptions underlying this approach. First, we assume that words whose relative frequency in the collection of topically related documents is much higher than their relative frequency in a general-topic collection are relevant for the topic. Second, we assume that event mentions consisting of topically relevant words are relevant for the topic.

Let D be the group of documents we wish to summarize and C a large general-topic corpus. We compute the informativeness score of an event e as:

$$S_I(e) = \sum_{w \in W(e)} \left(\frac{\#(w,D)}{\sum_{w' \in W(D)} \#(w',D)} - \frac{\#(w,C)}{\sum_{w' \in W(C)} \#(w',C)} \right)$$

where $W(C)$ is the set of all words in collection C, $W(e)$ is the set of all words constituting an event e (the event anchor and all the words from its arguments), and $\#(w,C)$ is the frequency of word w in collection C. We use the Google Books Ngrams corpus (Michel, Shen, Aiden, Veres, & Gray, 2011) as a large general-topic collection, C. Both D and C were lemmatized prior to computing the informativeness scores.

Although similar measures of sentence informativeness relying on relative word frequencies have been proposed in the literature (Conroy, Schlesinger, Goldstein, & Oleary, 2004; Lin & Hovy, 2000; Nenkova & Vanderwende, 2005), the key difference of our approach is that we focus on the informativeness of the individual event mentions rather than on all words of the text. Words that are not part of event mentions may be uninformative and may decrease the informativeness score of a sentence. We avoid this issue by considering only the words belonging to event mentions.

5.1.3. Smoothing over temporal structure

A unique feature of our model is that it takes into account the temporal structure of the document when estimating the relevance of events. We argue that, in addition to the intrinsic relevance originating from the relevance of its participants and overall informativeness, the relevance of an event can be estimated based on the event's temporal relations to other events. In other words, we assume that an event is more relevant for a topic if it is directly temporally related to other relevant events. For example, if an event e_1 happened after a relevant event e_2, we assume it is likely that e_2 is also relevant and that the relevance score of e_2 should be increased. To account for this, we rely on the structure of event graphs (i.e., temporal relations between event mentions) to smooth the relevance scores across the temporal structure. We employ PageRank (Page et al., 1998) as a smoothing (i.e., label propagation) algorithm.

PageRank was initially designed for ranking web pages by their relevance. The idea of PageRank is that a vertex in a graph should have a high score assigned to it if it has many neighbors with high scores. The score of a vertex should be even higher if its high-scoring neighbors do not have many other neighbors (i.e., their influence is instead distributed over a small number of neighbors). Let $\mathbf{W}$ be the row-normalized adjacency matrix of graph G. The algorithm iteratively computes the vector of vertex scores $\mathbf{a}$ in the following way:

$$\mathbf{a}^{(k)} = \alpha \mathbf{a}^{(k-1)} \mathbf{W} + (1-\alpha)\mathbf{s}$$

where α is the PageRank damping factor. Vector $\mathbf{s}$ models the normalized internal source of the score for all vertices, and its elements sum to 1. We use normalized participant relevance score (S_P) and event informativeness score (S_I) as the internal source of the score for graph vertices, i.e., s_i equals $S_P(e_i)/\sum_{j=1}^{n} S_P(e_j)$ when we smooth over participant relevance scores, and s_i equals $S_I(e_i)/\sum_{j=1}^{n} S_I(e_j)$ when we smooth over event informativeness scores. PageRank is executed twice for each document from the collection, first for the participant relevance scores and then for the informativeness scores.

5.1.4. Redundancy removal

Clustering of sentences for the purpose of removing redundancy is a common step in multi-document summarization (Christensen, Mausam, & Etzioni, 2013; Saggion & Gaizauskas, 2004). We employ a single-linkage agglomerative clustering algorithm (Gower & Ross, 1969) on top of a state-of-the-art model for measuring the semantic similarity of short texts (Šarić, Glavaš, Karan, Šnajder, & Bašić, 2012). Single-linkage clustering starts with each sentence constituting its own cluster, and then iteratively merging the two most similar clusters in each iteration. The similarity of clusters is defined as the maximum similarity between the pairs of sentences from both clusters. The clustering continues until the remaining cluster similarities fall below a predefined threshold. After obtaining the clusters, we select from each cluster the sentence with the highest event-based score as the most representative. Finally, we sort the selected representative sentences for each cluster with respect to their event-based scores in descending order and add them to the summary until reaching the predefined summary length.

5.2. Experimental setup

Because our event-centered summarization model performs extractive non-focused multi-document summarization of documents in English, we perform the evaluation on datasets that conform to the same criteria. The de facto standard is the datasets created within the Document Understanding Conference (DUC) shared evaluation tasks organized between 2001 and 2007. The non-focused multi-document summarization tasks for English are the DUC-2002 Extractive Summarization Task (Over & Liggett, 2002) (200-token summaries) and the DUC-2004 Task 2 (Over & Yen, 2004) (100-token summaries). The DUC-2002 dataset consists of 59 groups containing between 5 and 15 topically related news stories, while the DUC-2004 dataset consists of 50 groups, each with 10 news stories. Both datasets include gold standard summaries produced by humans for each group of topically related

Table 4
Summarization performance on DUC-2002 dataset (test). (Best results in each group are shown in bold.)

Model	ROUGE-1	ROUGE-2
DUC-2002 best (System 21)	0.395	**0.103**
DUC-2002 median (System 20)	0.365	0.086
DUC-2002 human	**0.418**	0.102
Event graph PR	0.397	0.099
Event graph EI	0.353	0.085
Event graph PR + EI	0.407	0.114
Event graph PR + EI + TS	**0.415**	**0.116**

Table 5
Summarization performance on DUC-2004 dataset. (Best results in each group are shown in bold.)

Model	ROUGE-1	ROUGE-2
DUC-2004 best (Conroy et al., 2004)	0.382	0.092
DUC-2004 baseline	0.324	0.064
DUC-2004 human	**0.440**	**0.134**
Event graph PR	0.385	0.096
Event graph EI	0.386	0.099
Event graph PR + EI	0.395	0.103
Event graph PR + EI + TS	**0.405**	**0.107**

documents. We evaluate the performance using ROUGE-1 and ROUGE-2 evaluation metrics (Lin, 2004), the most commonly used automated evaluation metrics for text summarization. Different evaluation metrics were used in DUC-2002. To allow for a comparison with these systems, we re-evaluated their summaries using ROUGE-1 and ROUGE-2.

Our summarization model contains two parameters that must be optimized: sentence clustering threshold t and the PageRank damping factor α. We split the DUC-2002 dataset into a development set (30 randomly selected groups) and a test set (the remaining 29 groups). We used the former set to optimize the parameters of the model (optimal values are $t = 0.3$ and $\alpha = 0.15$).

5.3. Results and discussion

The results for the DUC-2002 and DUC-2004 datasets are shown in Tables 4 and 5, respectively. We evaluate four variants of our event-centered summarization model: a model that uses only the participant relevance scores (PR), a model that uses only the event informativeness scores (EI), a model that combines these scores (PR + EI), and a complete model that also uses temporal smoothing (PR + EI + TS). We make a comparison against the best performing and median performing models from the respective shared tasks, as well as against human performance (i.e., the overlap between gold standard summaries).

All our models outperform (although not all significantly) the best performing models from the respective shared tasks. On the DUC-2002 test set, the model using only event participant relevance (EP) performs better than the model relying only on informativeness of events (IE), while on the DUC-2004 dataset these two models offer comparable performance. The combined EP + IE model outperforms both the EP and the IE model ($p < 0.05$; paired student's t-test), demonstrating the plausibility of combining the two relevance criteria. Finally, the full summarization model that additionally employs structural smoothing (PR + EI + TS) outperforms the EP + IE model (the performance difference is significant at $p < 0.05$ for ROUGE-1 and DUC-2004) and the best performing models from both shared tasks in terms of both ROUGE-1 and ROUGE-2 (difference is significant at $p < 0.05$). This demonstrates that using a temporal structure provided by an event graph is beneficial, and, in turn, validates our hypothesis that temporal relatedness can be used to propagate event relevance. Interestingly, the performance of the full event-centered summarization model comes close to human agreement (especially for DUC-2002), suggesting that there is little room for further improvement on these test collections.

Table 6 shows example summaries produced by the full summarization model for one DUC-2002 and one DUC-2004 group paired with the corresponding human-created summary (i.e., gold summary). A comparison of the automatically and human-created

Table 6
Example summaries (DUC-2002 and DUC-2004 groups).

Group	Extractive summary (Event graph PR + EI + TS)	Gold summary
DUC-2002 d068f (200 tokens)	*Checkpoint Charlie, the Berlin Wall border post that symbolized the Cold War, was hoisted into history today. Checkpoint Charlie went up in 1961 in the middle of the Friedrichstrasse boulevard after Communist East Germany erected the Berlin Wall to choke off a flood of refugees to the enclave of West Berlin. Berlin has been split into Soviet, U.S., British and France sectors and West Berlin remains technically separate from West Germany. With the wall being dismantled daily in anticipation of German unification, U.S. officials decided to remove Checkpoint Charlie with a grand flourish. The Soviet Union said today that a united Germany can join NATO after a five-year transition period during which all Soviet and U.S. troops would leave the country. Since East Germany overthrew its Communist government last fall and the German borders were opened, Checkpoint Charlie has become as superfluous as the crumbling Berlin Wall. Baker, Soviet Foreign Minister Eduard Shevardnadze and the foreign ministers from France, Britain and the two Germanys each heralded the end of the checkpoint as a symbol of change. The German and Allied foreign ministers are meeting in East Berlin to discuss the status of the city and a united Germany's role in world affairs.*	*Checkpoint Charlie, the most famous symbol of the Cold War was lifted into history today. At an invitation only ceremony attended by the foreign ministers of the four World War II allies, a U.S. army band played an old Berlin song as the checkpoint was hoisted onto a waiting flatbed truck, to be taken to a museum. Speaking at the ceremony, Soviet Foreign Minister Eduard Shevardnadze, the first Soviet foreign minister to visit Berlin, called for the removal of Berlin's status as an allied controlled city and the withdrawal of troops six months after the two German states unify. Secretary of State James Baker, citing Checkpoint Charlie's status as a Cold War symbol, expressed the hope that its removal would bury the conflicts it created. The decision to remove Checkpoint Charlie came as large sections of the wall were being dismantled and the opening of the border between the two Berlins made the checkpoint superfluous. Its removal was timed to coincide with a meeting of German and Allied foreign ministers in East Berlin to discuss the status of Berlin and a united Germany. Among the invited guests was West Germany's Chancellor Willy Brandt who was mayor of Berlin in the 1960s.*
DUC-2004 d30003t (100 tokens)	*The Spanish and British governments appeared Wednesday to be seeking shelter from the political storm brewing over the possible extradition of former Chilean dictator Augusto Pinochet to Spain. British police arrested Pinochet in his bed Friday at a private London hospital in response to a request from Spain, which wants to question Pinochet about allegations of murder during the decade after he seized power in 1973. A delegation of Chilean legislators lobbying against the possible extradition of Augusto Pinochet to Spain to face trial, warned Thursday that Chile was on the brink of political turmoil.*	*Former Chilean dictator Augusto Pinochet has been arrested in London at the request of the Spanish government. Pinochet, in London for back surgery, was arrested in his hospital room. Spain is seeking extradition of Pinochet from London to Spain to face charges of murder in the deaths of Spanish citizens in Chile under Pinochet's rule in the 1970s and 80s. The arrest raised confusion in the international community as the legality of the move is debated. Pinochet supporters say that Pinochet's arrest is illegal, claiming he has diplomatic immunity. The final outcome of the extradition request lies with the Spanish courts.*

summaries reveals that there is a considerable semantic overlap between the two.

6. Conclusion and future perspectives

With the amount of documents describing real-world events growing rapidly (e.g., news stories, intelligence reports, social media posts), it has become increasingly important to efficiently retrieve and concisely present event-oriented information. In text, real-world events are represented as event mentions, which describe the circumstances of an event. Individual event mentions are related to each other, giving rise to a structure of event mentions. Standard information retrieval models, which rely on shallow and unstructured document representations, fall short of capturing this semantics. Similarly, existing multi-document summarization models do not specifically account for the semantics of sentence-level events.

This article aimed to bridge this gap and addressed event-centered retrieval and summarization based on sentence-level event extraction. The contribution of this article is threefold. First, we proposed *event graphs*, a novel event-centered document representation that accounts for the semantics of events. An event graph filters and structures event-relevant information in the form of a directed graph in which vertices correspond to individual event mentions and edges correspond to the temporal relations between them. We have described a hybrid (rule-based and machine learning-based) approach for the robust extraction of event graphs from text. Second, building on event graphs, we proposed a novel model for event-centered information retrieval. The model employs graph kernels over event graphs as an expressive measure of similarity between queries and documents. On event-oriented document collections, our model significantly outperforms well-established event-agnostic retrieval models. The third contribution of this article is a novel event-centered multi-document summarization model. The model selects sentences for the summary based on the estimated relevance of event mentions computed using event graphs, and it achieves a significant performance gain over competitive text summarization methods. In sum, our results conclusively demonstrate that event graph representation can improve the performance on event-centered text retrieval tasks.

We believe that the presented models make an important contribution to information retrieval and multi-document summarization. Our work is among the first to show the usefulness of event-oriented information extraction techniques in an information retrieval context, and we expect more work on structured event-oriented document representations to follow. The results obtained on the multi-document summarization tasks prove that both events and temporal relations between them are useful for recognizing the most relevant information within a group of topically related documents. Furthermore, the results suggest that similar event-based approaches could be used to address closely related NLP tasks such as text simplification.

This work opens up a number of interesting research directions for the semantic processing of events from text. We can identify three main research lines. The first line of research is related to the extraction of event graphs and how its quality can be improved. We believe that improving the extraction of temporal relations between events would yield further performance improvements in event-centered retrieval and summarization. The second line of research would aim to enrich event graphs with other relations that can hold between events, such as causality, entailment, and spatiotemporal containment. Finally, it would be quite useful to consider applications of event graphs to other tasks (e.g., event template extraction, recognizing textual entailment) as well as other event-oriented domains (e.g., biography mining, tweets, police reports).

References

ACE. (2005). Evaluation of the detection and recognition of ACE: Entities, values, temporal expressions, relations, and events. <http://www.itl.nist.gov/iad/mig/tests/ace/2005/doc/ace05-evalplan.v3.pdf>.

Ahn, D. (2006). The stages of event extraction. In *Proceedings of COLING/ACL 2006 workshop on annotating and reasoning about time and events* (pp. 1–8).

Allan, J. (2002). *Topic detection and tracking: Event-based information organization* (Vol. 12). Springer.

Allen, J. R. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM, 26*(11), 832–843.

Amati, G. (2003). Probability models for information retrieval based on divergence from randomness (Ph.D. thesis). University of Glasgow.

Aone, C., & Ramos-Santacruz, M. (2000). REES: A large-scale relation and event extraction system. In *Proceedings of the Sixth Conference on Applied Natural Language Processing* (pp. 76–83). Association for Computational Linguistics.

Atkinson, J., & Munoz, R. (2013). Rhetorics-based multi-document summarization. *Expert Systems with Applications, 40*(11), 4346–4352.

Atkinson, M., & Van der Goot, E. (2009). Near real time information mining in multilingual news. In *Proceedings of the International Conference on WWW 2009* (pp. 1153–1154). ACM.

Baralis, E., Cagliero, L., Jabeen, S., Fiori, A., & Shah, S. (2013). Multi-document summarization based on the Yago ontology. *Expert Systems with Applications, 40*(17), 6976–6984.

Barzilay, R., McKeown, K. R., & Elhadad, M. (1999). Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (pp. 550–557). Association for Computational Linguistics.

Bejan, C., & Harabagiu, S. (2008). A linguistic resource for discovering event structures and resolving event coreference. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*.

Bethard, S. (2013). ClearTK-TimeML: A minimalist approach to TempEval 2013. In *Second Joint Conference on Lexical and Computational Semantics (∗SEM)* (Vol. 2).

Borgwardt, K. M. (2007). Graph kernels (Ph.D. thesis). Ludwig-Maximilians-Universität München.

Bramsen, P., Deshpande, P., Lee, Y. K., & Barzilay, R. (2006). Inducing temporal graphs. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06)* (pp. 189–198). Association for Computational Linguistics.

Büttcher, S., Clarke, C. L., Yeung, P. C., & Soboroff, I. (2007). Reliable information retrieval evaluation with incomplete and biased judgements. In *Proceedings of the ACM SIGIR* (pp. 63–70). ACM.

Canhasi, E., & Kononenko, I. (2014). Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization. *Expert Systems with Applications, 41*(2), 535–543.

Castells, P., Fernandez, M., & Vallet, D. (2007). An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Transactions on Knowledge and Data Engineering, 19*(2), 261–272.

Chang, A. X., & Manning, C. D. (2012). SUTIME: A library for recognizing and normalizing time expressions. In *Proceedings of Eighth International Conference on Language Resources and Evaluation (LREC 2012)*.

Christensen, J., Mausam, S. S., & Etzioni, O. (2013). Towards coherent multi-document summarization. In *Proceedings of NAACL-HLT* (pp. 1163–1173).

Conroy, J. M., Schlesinger, J. D., Goldstein, J., & Oleary, D. P. (2004). Left-brain/right-brain multi-document summarization. In *Proceedings of the Document Understanding Conference (DUC 2004)*.

Dang, H. T., & Owczarzak, K. (2008). Overview of the TAC 2008 update summarization task. In *Proceedings of Text Analysis Conference* (pp. 1–16).

Daniel, N., Radev, D., & Allison, T. (2003). Sub-event based multi-document summarization. In *Proceedings of the HLT-NAACL 03 Workshop on Text Summarization* (Vol. 5, pp. 9–16). Association for Computational Linguistics.

De Marneffe, M. C., MacCartney, B., & Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of Fifth International Conference on Language Resources and Evaluation (LREC 2006)* (Vol. 6, pp. 449–454).

Du, P., Guo, J., Zhang, J., & Cheng, X. (2010). Manifold ranking with sink points for update summarization. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (pp. 1757–1760). ACM.

Filatova, E., & Hatzivassiloglou, V. (2004). Event-based extractive summarization. In *Proceedings of ACL Workshop on Summarization* (Vol. 111).

Finkel, J. R., Grenager, T., & Manning, C. D. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL '05)* (pp. 363–370). Association for Computational Linguistics.

Gärtner, T., Flach, P., & Wrobel, S. (2003). On graph kernels: Hardness results and efficient alternatives. In *Learning theory and kernel machines* (pp. 129–143). Springer.

Ge, J., Huang, X., & Wu, L. (2003). Approaches to event-focused summarization based on named entities and query words. In *Proceedings of the 2003 Document Understanding Workshop*.

Gennari, S. P., Sloman, S. A., Malt, B. C., & Fitch, W. (2002). Motion events in language and cognition. *Cognition, 83*(1), 49–79.

Glavaš, G., & Šnajder, J. (2013). Recognizing identical events with graph kernels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp.797–803).

Glavaš, G., & Šnajder, J. (2013a). Exploring coreference uncertainty of generically extracted event mentions. In *Proceedings of the CICLing 2013* (pp. 408–422). Springer.

Gower, J. C., & Ross, G. (1969). Minimum spanning trees and single linkage cluster analysis. *Applied Statistics*, 54–64.

Grishman, R., & Sundheim, B. (1996). Message understanding conference-6: A Brief history. In *Proceedings of International Conference on Computational Linguistics (COLING 1996)* (Vol. 96, pp. 466–471).

Grover, C., Tobin, R., Alex, B., & Byrne, K. (2010). Edinburgh-LTG: TempEval-2 system description. In *Proceedings of the Fifth International Workshop on Semantic Evaluation (SemEval 2010)* (pp. 333–336). Association for Computational Linguistics.

Hammack, R., Imrich, W., & Klavžar, S. (2011). Handbook of product graphs. *Discrete mathematics and its applications*. 9781439813041. CRC Press.

Hatzivassiloglou, V., Gravano, L., & Maganti, A. (2000). An investigation of linguistic features and clustering algorithms for topical document clustering. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 224–231). ACM.

He, R., Qin, B., & Liu, T. (2012). A novel approach to update summarization using evolutionary manifold-ranking and spectral clustering. *Expert Systems with Applications, 39*(3), 2375–2384.

Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., & Cunningham, H., et al. (1998). University of sheffield: Description of the LaSIE-II system as used for MUC-7. In *Proceedings of the Seventh Message Understanding Conferences (MUC-7)*.

Kawahara, D., Shinzato, K., Shibata, T., & Kurohashi, S. (2013). Precise information retrieval exploiting predicate-argument structures. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)* (pp. 37–45).

Kingsbury, P., & Palmer, M. (2002). From treebank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC '02)* (pp. 1989–1993).

Kolomiyets, O., Bethard, S., & Moens, M. (2012). Extracting narrative timelines as temporal dependency structures. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*.

Kumaran, G., & Allan, J. (2004). Text classification and named entities for new event detection. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 297–304). ACM.

Li, W., Wu, M., Lu, Q., Xu, W., & Yuan, C. (2006). Extractive summarization using inter-and intra-event relevance. In *Proceedings of the 21st international conference on computational linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, association for computational linguistics* (pp. 369–376).

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop* (pp. 74–81).

Lin, C.-H., Yen, C.-W., Hong, J.-S., & Cruz-Lara, S., et al. (2007). Event-based textual document retrieval by using semantic role labeling and coreference resolution. In *IADIS international conference WWW/Internet 2007*.

Lin, C.-Y., & Hovy, E. (2000). The automated acquisition of topic signatures for text summarization. *Proceedings of the 18th conference on computational linguistics* (Vol. 1, pp. 495–501). Association for Computational Linguistics.

Llorens, H., Saquete, E., & Navarro, B. (2010). TIPSem (English and Spanish): Evaluating CRFs and semantic roles in TempEval-2. In *Proceedings of the Fifth International Workshop on Semantic Evaluation* (pp. 284–291). Association for Computational Linguistics.

Makkonen, J. (2003). Investigations on event evolution in TDT. In *Proceedings of the student research workshop at conference of the North American chapter of the association for computational linguistics on human language technology (NAACL-HLT '03)* (pp. 43–48). Association for Computational Linguistics.

Makkonen, J., Ahonen-Myka, H., & Salmenkivi, M. (2004). Simple semantics in topic detection and tracking. *Information Retrieval, 7*(3), 347–368.

Mayo, B. (1950). Events and language. *Analysis, 10*(5), 109–115.

Menchetti, S., Costa, F., & Frasconi, P. (2005). Weighted decomposition kernels. In *Proceedings of the 22nd international conference on machine learning* (pp. 585–592). ACM.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176.

Nallapati, R., Feng, A., Peng, F., & Allan, J. (2004). Event threading within news topics. In *Proceedings of the 13th ACM International Conference on Information and Knowledge Management* (pp. 446–453). ACM.

Nenkova, A., & Vanderwende, L. (2005). The impact of frequency on summarization. Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101.

Noh, T.-G., Park, S.-B., Yoon, H.-G., Lee, S.-J., & Park, S.-Y. (2009). An automatic translation of tags for multimedia contents using folksonomy networks. In *Proceedings of the ACM SIGIR 2009* (pp. 492–499). ACM.

Ouyang, Y., Li, W., Li, S., & Lu, Q. (2011). Applying regression models to query-focused multi-document summarization. *Information Processing & Management, 47*(2), 227–237.

Over, P., & Liggett, W. (2002). Introduction to DUC-2002: An intrinsic evaluation of generic news text summarization systems. In *Proceedings of Workshop on Automatic Summarization (DUC 2002)*.

Over, P., & Yen, J. (2004). Introduction to DUC-2004: An intrinsic evaluation of generic news text summarization systems. In *Proceedings of Workshop on Automatic Summarization (DUC 2004)*.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank citation ranking: Bringing order to the web. Tech. rep. Stanford Digital Library Technologies Project, Stanford.

Pai, M.-Y., Chen, M.-Y., Chu, H.-C., & Chen, Y.-M. (2013). Development of a semantic-based content mapping mechanism for information retrieval. *Expert Systems with Applications, 40*(7), 2447–2461.

Pan, Z., & Kosicki, G. M. (1993). Framing analysis: An approach to news discourse. *Political Communication, 10*(1), 55–75.

Ponte, J., & Croft, B. (1998). A language modeling approach to information retrieval. In *Proceedings of the ACM SIGIR* (pp. 275–281). ACM.

Pustejovsky, J. (1991). The syntax of event structure. *Cognition, 41*(1), 47–81.

Pustejovsky, J., Castano, J., Ingria, R., Sauri, R., Gaizauskas, R., Setzer, A., et al. (2003a). TimeML: Robust specification of event and temporal expressions in text. *New Directions in Question Answering, 2003*, 28–34.

Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., & Setzer, A., et al. (2003). The TimeBank corpus. In *Corpus linguistics* (Vol. 2003, p. 40).

Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information science, 27*(3), 129–146.

Saggion, H., & Gaizauskas, R. (2004). Multi-document summarization by cluster/profile relevance and redundancy removal. In *Proceedings of the Document Understanding Conference* (pp. 6–7).

Salton, G., Wong, A., & Yang, C. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613–620.

Šarić, F., Glavaš, G., Karan, M., Šnajder, J., & Bašić, B. D. (2012). TakeLab: Systems for measuring semantic text similarity. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), in conjunction with the First Joint Conference on Lexical and Computational Semantics (∗SEM 2012)*.

Sarkar, I. N. (2012). A vector space model approach to identify genetically related diseases. *Journal of the American Medical Informatics Association, 19*(2), 249–254.

Sbattella, L., & Tedesco, R. (2013). A novel semantic information retrieval system based on a three-level domain model. *Journal of Systems and Software, 86*(5), 1426–1452.

Suzuki, J. (2005). Kernels for structured data in natural language processing (Ph.D. thesis). Nara Institute of Science and Technology.

Suzuki, J., Isozaki, H., & Maeda, E. (2004). Convolution kernels with feature selection for natural language processing tasks. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL'04), Main Volume* (pp. 119–126). Association for Computational Linguistics.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research, 37*, 141–188.

UzZaman, N., & Allen, J. (2010). TRIPS and TRIOS system for TempEval-2: Extracting temporal information from text. In *Proceedings of the Fifth International Workshop on Semantic Evaluation* (pp. 276–283). Association for Computational Linguistics.

UzZaman, N., Llorens, H., Derczynski, L., Verhagen, M., Allen, J., & Pustejovsky, J. (2013). SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantics (∗SEM 2013)*. Association for Computational Linguistics.

Van Dijk, T. A. (1985). Structures of news in the press. *Discourse and Communication: New Approaches to the Analysis of Mass Media Discourse and Communication, 10*, 69.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., & Pustejovsky, J. (2007). SemEval-2007 task 15: TempEval temporal relation identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluation (SemEval 2007)* (pp. 75–80).

Verhagen, M., Sauri, R., Caselli, T., & Pustejovsky, J. (2010). SemEval-2010 task 13: TempEval-2. In *Proceedings of the Fifth International Workshop on Semantic Evaluation (SemEval 2010)* (pp. 57–62). Association for Computational Linguistics.

Wei, C. P., & Chang, Y. H. (2007). Discovering event evolution patterns from document sequences. *IEEE Transactions on Systems, Man and Cybernetics. Part A: Systems and Humans, 37*(2), 273–283.

Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL 1994)* (pp. 133–138). Association for Computational Linguistics.

Yang, Y., Carbonell, J. G., Brown, R. D., Pierce, T., Archibald, B. T., & Liu, X. (1999). Learning approaches for detecting and tracking news events. *Intelligent Systems and their Applications, 14*(4), 32–43.

Yang, C. C., Shi, X., & Wei, C. P. (2009). Discovering event evolution graphs from news corpora. *IEEE Transactions on Systems, Man and Cybernetics. Part A: Systems and Humans, 39*(4), 850–863.

Yan, R., Kong, L., Huang, C., Wan, X., Li, X., & Zhang, Y. (2011). Timeline generation through evolutionary trans-temporal summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 433–443). Association for Computational Linguistics.