International Journal on Digital Libraries (2020) 21:215–229
https://doi.org/10.1007/s00799-018-0246-x

Toward comprehensive event collections

Federico Nanni[1] · Simone Paolo Ponzetto[1] · Laura Dietz[2]

Received: 2 October 2017 / Revised: 2 June 2018 / Accepted: 6 June 2018 / Published online: 22 June 2018

Abstract
Web archives, such as the Internet Archive, preserve an unprecedented abundance of materials regarding major events and transformations in our society. In this paper, we present an approach for building event-centric sub-collections from such large archives, which includes not only the core documents related to the event itself but, even more importantly, documents describing related aspects (e.g., premises and consequences). This is achieved by identifying relevant concepts and entities from a knowledge base, and then detecting their mentions in documents, which are interpreted as indicators for relevance. We extensively evaluate our system on two diachronic corpora, the New York Times Corpus and the US Congressional Record; additionally, we test its performance on the TREC KBA Stream Corpus and on the TREC-CAR dataset, two publicly available large-scale web collections.

Keywords Event collections · Named events · Collection building · Entity query expansion · Web archives

1 Introduction

The World Wide Web provides the research community with an unprecedented abundance of primary sources for the diachronic tracking, examination and—ultimately—understanding of major events and transformations in our society. These materials have the potential of offering deeper understandings of phenomena such as the rise of Euroscepticism, the causes and consequences of the Arab Spring as well as the global shock provoked by the recent global Financial Crisis.

Given the known ephemerality of materials that are created and exist only in digital format [29,44], since the 1990s, public and private institutions have embraced the responsibility of preserving these resources for future studies [18]. While in the last 20 years web archiving initiatives such as the Internet Archive [33] and the UK Web Archive [8] have made a lot of progress in terms of preservation, these collections are now so vast that—in the rare cases when they are fully available for research [23]—it is infeasible for scholars to conduct close reading analyses of specific topics [57]. In order to address this issue and for sustaining the use of the collected resources in humanities and social science research, a common approach currently adopted by web archive institutions is to offer manually curated topic specific collections, generated through a very time-consuming process.

The task To overcome these limitations, in our research[1] we focus on the task of automatically building event collections from large corpora of previously harvested documents (such as news, transcript of political speeches, web pages or social media posts). Given a specific named event (e.g., the 2004 Ukraine Orange Revolution) in the form of a URI of a Wikipedia page, the goal is to select a set of relevant documents that will be further analyzed, for example, by a historian through close reading. Therefore, the collection needs to be high in precision while maintaining breadth and comprehensiveness, i.e., to include information on premises and consequences. While the restriction to events on Wikipedia may seem like a limitation of applicability, we envision historians extending Wikipedia with domain-specific knowledge, in order to adopt our solution for particular events (as remarked in [43,50]).

✉ Federico Nanni
federico@informatik.uni-mannheim.de

Simone Paolo Ponzetto
simone@informatik.uni-mannheim.de

Laura Dietz
dietz@cs.unh.edu

[1] Data and Web Science Group, University of Mannheim, Mannheim, Germany

[2] Department of Computer Science, University of New Hampshire, Durham, USA

[1] This article builds upon and expands our previous work [41].

Our contribution In order to achieve this goal, in our previous work [41] we have proposed a learning-to-rank (L2R) approach and an accompanying system for creating event collections suitable for retrospective historic analyses. Our method selects not only the core documents related to the event itself, but most importantly documents which describe related aspects, such as premises and consequences. It does so through the use of relevant concepts and entities, collected from a knowledge base, whose presence in documents is interpreted as one of many indicators of relevance.

In this extension of our previous work, we provide a more throughout presentation of our approach and experimental setting and expand the evaluations of entity, passage, and document selection by adopting what we have defined as entity and event aspects [42]. We additionally offer an in-depth analysis of the trade-off between the amount of training data and the performance of our L2R system on two large diachronic corpora: (a) news (New York Times Corpus: 1987–2007) and (b) transcript of political speeches (US Congressional Record: 1989–2016). In order to compare the performance of our approach across both datasets, we consider a set of 44 events among general elections, political crises and civil wars—assessed and evaluated on both corpora by domain experts.[2] To additionally benchmark our approach, we (c) further include a third experiment on a large (10TB) and publicly available web archive, namely the TREC KBA Stream corpus,[3] which comprises both news and social media posts, collected between 2011 and 2013 and (d) have added for this journal extension a case study on enrichment of Wikipedia articles of events using the new TREC-CAR dataset.[4]

Outline In Sect. 2, we offer an overview on the task of event collection building, while in Sect. 3 we present the works that are most related to our study. In Sect. 4, we describe each component of our system. Following, we introduce in Sect. 5 the datasets for evaluation and in Sect. 6 provide in-depth quantitative performance results of each step of our work. A discussion on the advantages and limits of our system is finally presented in Sect. 7.

2 Background: building event collections

The task of building event collections from large corpora, which we have been tackling in our research [41], has recently attracted the interest of the digital library community [19] as it is closely related, but differs in scope, to the task of event harvesting. Event harvesting focuses on collecting documents related to a new topic from the live web, with the primary goal of preservation [36].[5] The focus is on obtaining a high-recall set of documents for further filtering at a later stage. In contrast, the task of building event collections starts from a previously harvested archive and aims at retrospectively selecting the documents related to a given event. An advantage of the retrospective approach is that we can leverage information from knowledge bases, such as Wikipedia, when building the collection. As event harvesting operates under real-time constraints, this is often not possible during the harvesting stage.

Manually curated event collections In the recent years, web archive institutions started to offer manually curated event collections. On Archive-It, for example, the Internet Archive presents a few collections regarding large-scale events such as the Boston Marathon shooting, the Black Lives Matter movement and the Charlie Hebdo terrorist attack [49,55]. These collections are created and curated by "the Archive-It team in conjunction with curators and subject matter experts from institutions around the world."[6] The same approach has been employed by public institutions.[7].

Current limitations The collections created with this manual approach have limitations: (a) They are small in number and in size, because manual selection is an extremely time-consuming process.[8] For example, Archive-It offers only 25 collections: These are focused on a few recent global events (e.g., the Ukraine War), but many others are missing (e.g., the Refugee Crisis); (b) Additionally, the selection process is not completely transparent, with selection guidelines (i.e., what to include and what not) not being publicly available.

Pros and cons of Event-name filtering Instead of creating these collections manually, automatic methods can also be adopted. For example, a document filtering approach which selects only the documents that mention the name of the event has been employed by researchers for the Temporal Summarization Task organized by the Text Retrieval Conference (TREC) [5].

While this approach was designed to obtain an initial high-recall collection (i.e., a superset of relevant documents), we argue in this paper that the resulting corpus is still not comprehensive enough for researchers in the humanities and the social sciences. If we are in fact to build a collection for the 2004 Ukraine Orange Revolution and only retrieve documents that precisely mention the name of the event, we will

[2] All gold standards are available at: http://federiconanni.com/event-collections/.

[3] http://trec-kba.org/kba-stream-corpus-2014.shtml.

[4] http://trec-car.cs.unh.edu/.

[5] See, for example, Nick Ruest collection of the Bataclan Attack: http://ruebot.net/post/look-14939154-paris-bataclan-parisattacks-porteouverte-tweets.

[6] More info here: https://archive-it.org/organizations/89.

[7] For example, the UK Web Archive: https://www.webarchive.org.uk/ukwa/collection.

[8] Size of the collections varies, spanning from 18 documents to more than 6000, depending on the topic.

miss materials that connect the origin of the revolution to the previous controversial presidential election in the country. And the same issue will emerge when studying the first Algerian democratic elections since independence (1990), which are a premise of the following Algerian civil war, or even when investigating the economic crisis behind Fujimori's *auto-golpe* in Peru, 1992. In this last case, the documents that discuss to adopt austerity measures will be not be included in the collection.

3 Related work

The task we address in our research is to create comprehensive event collections by retrieving materials from large datasets (e.g., newspaper corpora, web archives), in order to support research in the humanities and the social sciences. The methodological part of this work is therefore set at the intersection of three research areas: Firstly, it is related to the automatic retrieval of textual information concerning an event from a collection of documents; Additionally, our work focuses on taking advantages of the existing relations (expressed in knowledge bases) between named events and other named entities; finally, our work is connected to the use of entities and language models to expand event-related queries.

Events in NLP and IR For the last 20 years, the Natural Language Processing (NLP) and Information Retrieval (IR) communities have been working on the detection, extraction and tracking of events. The foundations for collection building and harvesting go back to a classic IR task called document filtering [30]. In document filtering, a stream of documents is to be filtered to only the ones about a given information need. More recently, the TREC Knowledge Base Acceleration track began to study how to track people and organizations in a diachronic collection, by building language models of entities that change over time [11].

Early efforts on tracking events in a stream of news were made in the Topic Detection and Tracking Task (TDT) at the TREC [2]; and related to it, the First-Story Detection Task was focused on retrieving the first document related to a new event in a stream of news [3]. In more recent years, the TREC Temporal Summarization track has aimed to provide introspective passage summarizations of an event as it is unfolding [26].

In contrast, the NLP community has mainly focused on the extraction of fine-grained events, which constitute n-ary relations between entities, such as time and location. For example, an event extracted from the sentence "Angela Merkel went to D.C. in August" connects the entities Angela Merkel, D.C. and August with the predicate "went to." During the last decade, thanks to the efforts in developing annotation guidelines, conducting evaluation campaigns[9] and organizing specific workshops,[10] the task of event extraction has attracted much attention in the field. The approaches developed in this area are often based on a combination of different machine learning models which employ morphosyntactic as well as temporal features [9,17].

Given the importance of events as a topic of study in historical research, [54] have recently studied whether the efforts of the NLP community on event extraction could be beneficial for supporting such studies (for example, via the creation of event collections). Interestingly, they pointed out how, among seventy-four interviewed historians, almost all of them agreed in recognizing 'historical events' in the form of coarse-grained named events (i.e., events which have a name and appear in a knowledge base such as DBpedia [7], for example the Korean War), while results were way less consistent for what concerned fine-grained (especially single-token) events, which are the typical output of the event extraction task. It is also interesting to note that when event collections are created by public and private archival institutions, they are also generally built around named events (e.g., the Charlie Hebdo Shooting[11]). For these reasons, in our work we focus on building collections for named events.

The paper that is closer to our work is by [19]. In this recent study, the authors generate event collections from the German Web Archive through the use of *1)* hyper-link analysis and *2)* lexical similarity (TF-IDF). While their methodology is not applicable to our experimental setting, as the corpora used in this work are not hyperlinked, in the evaluation we study the advantages of adopting semantic features (i.e., word embeddings, such as state-of-the-art GloVe [45]) over lexical features (i.e., TF-IDF) for the task of building event collections.

Events and entities The importance of employing geographical [15] and temporal [25] information in order to gain a better understanding of social phenomena through language is a relevant topic in NLP. A large amount of work focuses on detecting stories (such as events) in documents [4], combining historical events with information from social media [20], generating event digests from Wikipedia [38] and building time-aware exploratory search systems [12,53] (often considering the name of the event as the query [37]). The task of extracting important events adopting named entities has been recently addressed by [1] and by [21]. Entities have also been used to study the general perception of society toward past events [6], and their movements have been mapped extracting information from Wikipedia [34,51]. [28] employ named entities to extract yet unknown named events for knowledge

[9] https://www.ldc.upenn.edu/collaborations/past-projects/ace; http://www.timeml.org/tempeval/.

[10] For example: https://sites.google.com/site/cfpwsevents/.

[11] https://archive-it.org/collections/5541.

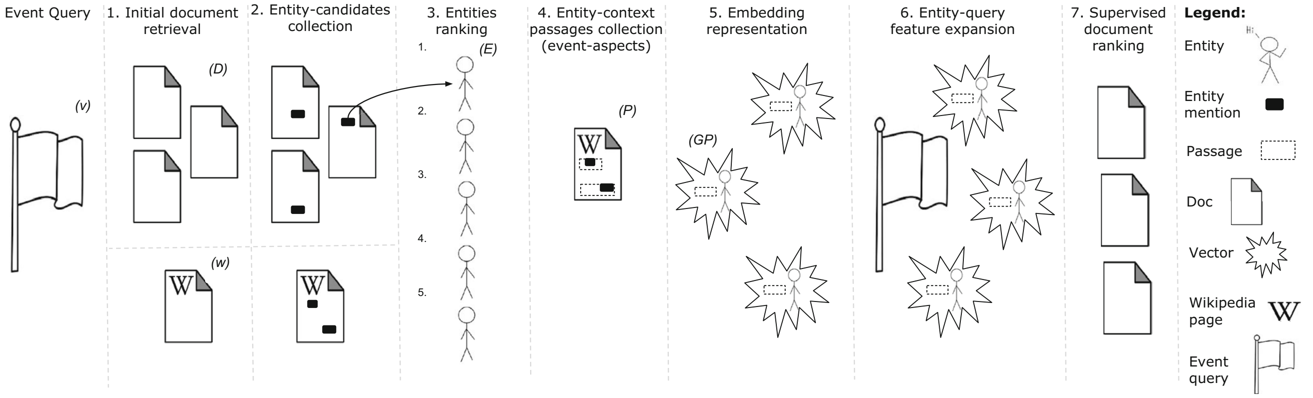
Event Query
1. Initial document retrieval
2. Entity-candidates collection
3. Entities ranking
4. Entity-context passages collection (event-aspects)
5. Embedding representation
6. Entity-query feature expansion
7. Supervised document ranking
Legend:
(v)
(D)
(w)
(E)
1.
2.
3.
4.
5.
(P)
(GP)
Entity
Entity mention
Passage
Doc
Vector
Wikipedia page
Event query

Fig. 1 Pipeline schema of our system

base population. Our work aims in a different direction, by first extracting entities which are related to a known named event and then projecting them back in time, in order to retrieve a comprehensive set of relevant documents for retrospective analysis.

Entity-query feature expansion Our approach is related to recent advances in information retrieval to exploit knowledge graphs. This includes approaches that tap into linguistic knowledge bases such as WordNet [27,35], as well as retrieval and scoring methods that use entity link annotations (i.e., annotations connecting the mentions of entities to knowledge base entries) for term matching and query expansion [10,22,47]. Combinations of knowledge base retrieval and entity linking methods have been studied for web search queries both for entity ranking tasks [46,52] as well as document ranking tasks [13,32,58]. Our work builds on these ideas for the purpose of creating event collections.

4 System's overview

Our system for building event collections consists of seven components, as depicted in Fig. 1. The user selects a named event v of interest, such as the 2004 Orange Revolution and a specific document collection C, for example the New York Times Corpus. As remarked above, the named event is expected to be an entity in DBpedia [7]. That is, given the alignment between DBpedia and Wikipedia, v corresponds to the title of a Wikipedia page $w \in W$.

Phase 1: initial document retrieval Retrieve an initial set of documents D from C using the name of the event v as a query q and collect all the documents with a mention of the Event-name, so for example "Orange Revolution."

Phase 2: entity candidate collection Extract a set of potentially relevant entities E, such as for example Yulia Tymoschenko and Viktor Yushchenko for the Orange Revolution, from two resources: the pool of relevant documents D and the Wikipedia page w corresponding to the event. Entities from D are extracted using the entity linker TagMe2 [16]; in particular, we collect all entities in the surrounding context of the event mentions using a context window of three sentences. Entities from w are collected following all Wikipedia outlinks. This approach is inspired by work on entity query feature expansion [13,32].

Phase 3: entity ranking Rank entities from E by relevance to the event. Since entities and named events refer to Wikipedia pages, which, in turn, are aligned with Wikipedia-centric knowledge base entries like, for instance, DBpedia entities, we can leverage knowledge base embeddings for this purpose. That is, we compute entity-event relatedness as the cosine similarity (cs) of entity and event vector representations using the pre-computed RDF graph embedding representations provided by RDF2Vec [48] (using a 500-dimensional vector space), and rank based on that (as presented in Sect. 6.2).

Phase 4: entity-context passage collection For each entity $e \in E$, collect a supporting text passage $p(e)$ presenting the relation between the entity and the event, for instance, the fact that Yulia Tymoshenko co-led the Orange Revolution and was the first woman appointed Prime Minister of Ukraine. To do so, we retrieve, from the Wikipedia page w of the event, the content of the section that is most related to the entity e. Relatedness is computed as the cosine similarity of entity-name and event-section vector representations using the state-of-the-art pre-computed GloVe word embeddings (300d) [45] representations. We call this content an *event aspect* and refer to it as **EvAsp** in the rest of the paper.

Phase 5: embedding representation Project entities E and contextual passages P into an embedding space, as to obtain their latent feature vectors: GE and GP, respectively. These are obtained by computing the element-wise average of the embeddings for entities in E and passages in P, respectively. Let S_p be, for example, the set of unique words of a passage

$p(e) \in P$, namely an event aspect of entity e. The embedding of $p(e)$ $(gp(e) \in GP)$ is then computed as:

$$gp(e) = \frac{1}{N} \sum_{w \in S_p} \mathrm{freq}(w) \cdot \mathbf{v}_w \quad (1)$$

where freq(w) is the frequency of occurrence of word w in passage p, $\mathbf{v}_w$ is the embedding vector for word w, and N is the total number of words in p. The same formula is applied to obtain entity embeddings $ge \in GE$. Again, we use the state-of-the-art pre-computed GloVe word embeddings (300d) [45].

Phase 6: entity-query feature expansion Extend the initial event query q with the following vector space expansion models.

- *Place* Expansion with the location entity L only (e.g., Kiev, which is retrieved from the knowledge base) using a TF-IDF vector. We argue that, in specific cases, the location is already a precise indicator for retrieving relevant documents.
- *Ent-TFIDF* Expansion with the top-10 related entities from the set E, ranked as described in Phase 3, using TF-IDF.
- *EvAsp-TFIDF* Expansion with words from the contextual passages P (as collected in Phase 4 and defined as event aspects) of the top-10 related entities, using a TF-IDF vector representation.
- *Ent-GloVe* Expansion with the top-10 related entities from the set E, ranked as described in Phase 3, using GloVe vector representations.
- *EvAsp-GloVe/Our-light* Expansion with words from the contextual passages P (as collected in Phase 4 and defined as event aspects) of the top-10 related entities, using GloVe vector representations. It is our championed method, which we refer to as **Our-light** in the rest of this paper.

Phase 7a: query processing and ranking Given the named event v (i.e., an entity) and an expansion set consisting of either additional entities (Place, Ent-TFIDF, Ent-GloVe) or words from event aspects (EvAsp-TFIDF, EvAsp-GloVe), we process entities and entity aspects as follows, in order to normalize them to a query q. We first interpret entities as words—e.g., the entity Yulia Tymoshenko is represented as the words "Yulia" and "Tymoshenko." We then represent each word—either from the named event, or the entity or event aspect expansion—as a vector, and build a query vector as the element-wise sum of all word vectors. We study two variations of vector space models: TF-IDF (logarithmic, L2-normalized variant) over the corpus vocabulary and the GloVe word embedding. The result set is then computed by ranking documents in C according to the cosine similarity (cs) of query and document vector. Features are summarized in Table 1.

Table 1 Features used in Our-full approach in order to rank documents

Type of query	Feature
Event-name	TF-IDF (cs)
Place	TF-IDF (cs)
Entities	TF-IDF (cs), GloVe (cs)
Event aspects	TF-IDF (cs), GloVe (cs)

Phase 7b: supervised document ranking/Our-full We combine the ranking-score of the different methods studied in Phase 7a (see Table 1) as features in a list-wise learning-to-rank (L2R) setting [31] implemented in RankLib[12] for producing a final ranking of relevant documents, starting from all documents in C. The weight parameter is learned by optimizing for the mean average precision (MAP) of the ranking using coordinate ascent. L2R will learn a weighted feature combination to achieve the best possible ranking on the training set. We study feature sets for their merit by applying L2R on held-out test data using fivefold cross-validation.

In the experimental section, we provide evidence on the usefulness of the different features for retrieving a more comprehensive set of documents, which cover, for instance, also the outcome of the 2004 Ukrainian Presidential Election as a premise of the Orange Revolution. We refer to this supervised approach as **Our-full**.

5 Experimental setup

5.1 Datasets

We test our system on four collections. Their differences (news vs. political speeches vs. online content, small-scale vs. large-scale datasets) permit us to assess the performance of our approach in various research contexts and with different types of events.

NYT Corpus The New York Times Corpus comprises over 1.8 million articles published between 1987 and 2007.[13]

USC Corpus The US Congressional Records is a collection of all proceedings of the US Congress. We collected this corpus from THOMAS at the beginning of 2016, when the original website was still available online.[14] The obtained

[12] https://sourceforge.net/p/lemur/wiki/RankLib/.

[13] https://catalog.ldc.upenn.edu/ldc2008t19.

[14] THOMAS has been a digital collection directed by the Library of Congress. It offered, among other materials, the official record of proceedings and debate since the 101th Congress (1989–1990). In 2016,

corpus spans for more than 26 years (1989–2016). For each day, we collected transcriptions of all statements given on the Senate and the House floor, plus the related Extensions of remarks. This collection sums up to over 1.2 million documents.

KBA Corpus To evaluate in a large-scale experimental setting, we consider the 2014 TREC KBA Stream Corpus, a large web archive collection (10TB) of news, social media posts, forums and scientific publications collected from the web between October 2011 and January 2013.

TREC-CAR As an additional evaluation regarding Wikipedia enrichment, we adopt the recently introduced Complex Answer Retrieval dataset,[15] where the organizers processed the English Wikipedia, associating each paragraph in each page with a related query (e.g., the page name, the section-heading).

5.2 Types of events

Some types of events are easier to track in text compared to others, for example, pre-planned events which had an established name before happening, such as referendums (e.g., Brexit),[16] sport events (the 2016 Olympic Games) or concerts (Eurovision 2016), as well as events that suddenly happen without any direct premise, like natural disasters (the Fukushima nuclear disaster) or terrorist attacks (the Bataclan Attack). As a matter of fact, these events can be simply tracked in text by searching for mentions of the Event-name (example: retrieve all documents that mention "Brexit"). However, while this approach could produce satisfying event collections for certain types of events (or for certain kinds of tasks, such as event summarization), we argue in this paper that it provides unsatisfying results when trying to collect materials for obtaining a comprehensive overview of complex events that grow and evolve during time, such as political crises, protests as well as civil wars. In order to assess the correctness of our assumption, we consider the following different types of events.

Unexpected elections The first type is what we call here "unexpected political elections." An unexpected political election could be due to the beginning of a democratic transition[17] as well as the result of a political crisis.[18] We identified 15 unexpected elections, which took place between 1989 and 2007 using the National Elections Across Democracy and Autocracy data-set (NELDA) [24].[19] In particular we considered elections flagged with the variables NELDA 2 ("Were these the first multi-party elections?") or NELDA 6 ("If regular, were these elections early or late relative to the date they were supposed to be held per established procedure?").[20]

Political crises The second type is political crises. While these events are easy to track in text through string matching of the Event-name (e.g., the Cassette Scandal, which happened in Ukraine in 2000), we assume that their retrieval in documents becomes more complex when they are in their early stages and the name is still not established or the crisis has not yet emerged. We identified 15 political crises, combining information from the NELDA dataset with a set of Wikipedia categories on the topic.[21]

Civil wars The third type is civil wars. While tracking events such as wars could be done using a combination of specific keywords (e.g., "war," "invasion," "battle") and the name of the involved countries, internal wars (such as the conflicts that brought to the breakup of Yugoslavia) are way more complex to track and often arise as a consequence of previous long-term internal political tensions. Therefore, we argue that these instabilities cannot be easily captured by simply searching for documents that mention the name of the event (e.g., Bosnian War). We identify 14 civil wars, combining information from the NELDA dataset with a set of Wikipedia categories on the topic.[22]

6 Experimental evaluation

In this section, we first evaluate the quality of the approach we adopt for collecting and ranking entities that are related to an event. Next, we establish the quality of the retrieved contextual passages. Finally, we test the performance of our system for ranking documents that are relevant to a specific event, in particular by comparing the results with the most-employed automatic method for the task: retrieving documents that contain mentions of the Event-name.

6.1 Collecting entities

As our approach distinguishes collecting (see Phase 2) and ranking (see Phase 3) entities, we study the performance

THOMAS has been completely substituted with Congress.gov, which provides full-text access to daily congressional record issues dating from 1995 (beginning with the 104th Congress).

[15] http://trec-car.cs.unh.edu/.

[16] As also remarked in [19].

[17] Cf. e.g., the first multi-party election in Algeria, 1991.

[18] See for example the Italian general election in 1996.

[19] http://www.nelda.co/.

[20] A list of all events examined in our work is available here: https://federiconanni.com/event-collections/.

[21] https://en.wikipedia.org/wiki/Category:Protests; https://en.wikipedia.org/wiki/Category:Economic_crises; https://en.wikipedia.org/wiki/Category:Government_crises.

[22] https://en.wikipedia.org/wiki/Category:20th-century_conflicts_by_year; https://en.wikipedia.org/wiki/Category:Civil_wars.

of each component in isolation. Given a named event, such as an election, an internal conflict or an anti-establishment protest, we compare the solution we decided to adopt with other approaches.

6.1.1 Gold standard

For every event, each approach presents a pool of candidate entities. We consider, in this step of the work, a subset of 20 events. The relevance of each entity to each event has been manually assessed by two domain experts on a binary scale. The obtained result, which is composed of 830 annotated entity-event pairs (484 relevant and 346 not relevant), extends the gold standard of entity-event relatedness assessments we created for a previous work [40].

6.1.2 Methods for collecting entities

In Phase 2, our system retrieves a pool of potentially relevant entities (a) from initially collected relevant documents and (b) by following the outlinks in the Wikipedia page of the event. We call our method **Cont + Out**. We study the performance of our approach and compare it with (a) the performance of each of its components in isolation (**Context** and **Outlinks**) and (b) the following baselines:
Info-box For each event, all entities that appear in the Info-Box of the Wikipedia article of the event are selected.
NELDA The NELDA dataset includes a manually selected list of related entities for specific political scenarios (e.g., political leader(s) of the country, before and after an election). We include this as a manual (upperbound) reference baseline.

6.1.3 Results on entity collection

For each event, the different approaches for collecting potentially relevant entities present a set of candidates. Given our gold standard annotations, in Table 2 we report precision, recall and F1-Score. We can notice that a political science dataset such as NELDA is limited for our goal, as it provides only a small number of relevant entities. Other approaches, such as collecting entities from info-boxes and contextual passages, have similar drawbacks (i.e., extracting too few or many unrelated entities, while in both cases missing a few central ones). In particular, when analyzing the results obtained by collecting contextual entities, we noticed that—from time to time—the event is mentioned out of context, for example as part of a comparison, and therefore the collected entities are not related.
Take-away To conclude, while using Wikipedia Outlinks leads to good results, the best performance is obtained when creating a pool of entities by combining candidates collected from Wikipedia and candidates retrieved from contextual passages. This finding is in line with experiments of [13]. Consequently, we use this approach for Phase 2 of our system.

Table 2 Precision, recall and F1-Score on entity collection

Method	Precision	Recall	F1
NELDA	**1.00 ± 0.00**	0.13 ± 0.02	0.23 ± 0.02
Info-box	0.88 ± 0.03	0.27 ± 0.05	0.41± 0.05
Context	0.52 ± 0.04	0.60 ± 0.05	0.55 ± 0.05
Outlinks	0.89 ± 0.03	0.53 ± 0.05	0.66 ± 0.05
Cont + Out	0.74 ± 0.04	**1.00 ± 0.00**	**0.85 ± 0.05**

Bold indicates the higher number in each column

6.2 Ranking entities

As a second step, we study the performance of the entity ranking method we employed (Phase 3), in comparison with other approaches, both graph and content based.

6.2.1 Methods for ranking entities

In Phase 3, we rank entities by computing the cosine similarity between the RDF embedding representation of entities and events; we report the results computed on the NYT Corpus.
RDF2Vec This approach establishes semantic similarities between event and entities by ranking entities, with respect to the event, by the cosine similarity of their RDF graph embedding representations [48]. In our work, we use entity embeddings with 500 dimensions computed by Ristoski and Paulheim and we consider as an initial pool of entities to rank all entities collected by the other baselines.

We study the performance of RDF2Vec in comparison with the following methods:
ContFreq Rank the set of entities by their raw frequency of occurrence in relevant context. We assume that important entities appear often in the context of an event mention.[23]
CheapEntRel Use a rank-based aggregation ($\sum \frac{1}{r_E}$) of the following four rankings, adopting a variation of linked-based TF-IDF (log variant with L2 normalization) and employing document frequency statistics from DBpedia (Version 04-2015)):

- Rank entities linked in the event's article by TD-IDF (outlink).
- Rank entities by how often they link back to the event's article (backlink).
- Rank entities by the ratio of outlink frequencies divided by backlink frequency.

[23] We also tested TF-IDF weighted frequency, but we did not obtain any significant improvement over raw frequency.

– Rank entities according to the **ContFreq** baseline.

This method was used in our previous study [40] and is inspired by the work of [56].

Event/entity aspects We additionally experiment with the textual content of the Wikipedia pages of the event and the entity, in order to measure their relatedness. Following the intuition that only a few "aspects" (i.e., sections) of an event will be related to an entity and vice versa,[24] we rank entities in two different fashions. First, by computing the cosine similarity between the vector representation of the entity-name and all sections on the Wikipedia page of the event; we call this approach **EntName-EvAsp**, as it employs event aspects. Next, we adopt the vector representation of the Event-name and all sections on the Wikipedia page of the entity, which we call **EntAsp-EvName**. We test and report results using both **TF-IDF** and **GloVe** vector representations.

6.2.2 Results on entity ranking

Using the same gold standard introduced in the previous evaluation, we study the quality of the rankings using mean average precision (MAP) metric and by reporting the micro-averaged precision at 10; the results are presented in Table 3. We can notice how ranking contextual entities by their frequency is not a good approach, especially because it happens that related entities simply do not appear in the close proximity of the event mention (but they are mentioned in other parts of the same document). Comparing the cheap entity-relatedness method [40] and RDF2Vec shows that while our low-cost approach yields to good results, RDF2Vec clearly outperforms it. In addition, while one of the entity aspect approaches (EntAsp-EvName-GloVe) significantly outperforms RDF2Vec in MAP, all of them show lower performance for what concerns P@10.

Take-away As we want to collect a candidate set of relevant entities, we employ RDF2Vec for Phase 3 of our system and collect the top 10 entities retrieved by using this approach.

6.3 Collecting contextual passages

The next step of our work is to collect passages where each relevant entity $e \in E$ is presented in the context of the named event v (see the use of event aspects in Phase 4).

6.3.1 Gold standard

Using a subset of 312 relevant entities, for each entity we display all passages to two domain experts and ask whether each of these passages describes the relationship between the entity and the event. The obtained results comprise 751 annotated passages (570 relevant, 181 not relevant) and extend the gold standard of entity-passage relatedness assessments we created for a previous work [40].

Table 3 MAP and P@10 on entity ranking

Method	MAP	P@10
ContFreq	0.22 ± 0.03	0.48 ± 0.05
CheapEntRel	0.51 ± 0.05	0.62 ± 0.05
EntName-EvAsp-TFIDF	0.64 ± 0.04	0.29 ± 0.05
EntName-EvAsp-GloVe	0.66 ± 0.04	0.43 ± 0.06
EntAsp-EvName-TFIDF	0.62 ± 0.04	0.30 ± 0.05
EntAsp-EvName-GloVe	**0.80** ± **0.03**	0.56 ± 0.06
RDF2Vec	0.65 ± 0.05	**0.74** ± **0.05**

Bold indicates the higher number in each column

6.3.2 Methods for collecting entity contexts

We compare the quality of our approach (**EvAsp**, Phase 4) with the following baselines:

Wiki-intro Retrieve the first sentences of the Wikipedia page of the entity. In case the entity is highly related to the event, we assume this passage will elaborate on their relation.

Contextual passages Extract contextual passages from documents that mention the Event-name. We extract passages both from NYT articles and from speeches in the USC Corpus and report their effect separately (**NYT-Pass** and **USC-Pass** in Table 4).

6.3.3 Results on collecting entity contexts

For each entity, the different approaches for collecting potentially relevant passages present a candidate. Using our gold standard annotations, we report in Table 4 the precision, recall and F1-Score of the different approaches. As can be seen, adopting a baseline such as Wiki-Intro provides correct passages for less then half the entities. Additionally, while collecting passages from relevant documents is a good approach, only a small set of relevant entities can be captured in the proximity of the event mention. (The same issue emerges when ranking entities from contextual passages.) Another common issue arising in USC speeches is that when the event is mentioned as an aside, such as an enumeration, the context is not relevant for our task.

Take-away Collecting passages as the most related section (i.e., aspect) of the Wikipedia page of the event provides the overall best performing approach for this task, and therefore, we use this approach in our system.

[24] For example, the youth organization PORA is related to the aspects *Protests* and *Internet usage* of the event Orange Revolution and less to its *Causes*.

Table 4 Precision, recall and F1-Score on passage selection

Method	Precision	Recall	F1
Wiki-Intro	0.45 ± 0.03	**1.00 ± 0.00**	0.62 ± 0.03
NYT-Pass	**0.99 ± 0.03**	0.36 ± 0.03	0.53 ± 0.03
USC-Pass	0.92 ± 0.03	0.19 ± 0.03	0.31 ± 0.03
Ev-Asp	**0.99 ± 0.03**	0.81 ± 0.03	**0.89 ± 0.03**

Bold indicates the higher number in each column

Table 5 Statistics of the gold standards

Dataset	TOT	Rel	Not-Rel
NYT Corpus	1836	634	1202
USC Corpus	1861	612	1249
All	**3697**	**1246**	**2451**

Bold indicates the higher number in each column

6.4 Retrieving relevant documents

The final step of our evaluation is assessing the quality of our entire system for the task of retrieving documents related to an event. We present the performance of both our full pipeline (**Our-full**) and of its *light* version (**Our-light**), where full includes several methods with learning to rank and light includes the best single unsupervised method.

6.4.1 Gold standard

For each event, we consider an initial pool of documents in each corpus as a starting sub-corpus. These documents have been selected following these two premises: (a) they are published maximum 18 months before or after the event (i.e., within a 3-year window); (b) they contain the mention of the location where the event happened (e.g., the country or the city, depending on the event) as a very coarse-grained initial filter. On the obtained sub-corpus, we compare the performance quality of our approach for ranking relevant documents to several baselines.

Annotations We follow a pooled evaluation approach, which is common in the TREC community. For each of the 44 events, we use all baselines and systems to rank documents and then retain the top 15 documents in each ranking for manual assessment. Given the complexity of some of the studied events, instead of employing crow-sourcing, we hired two domain experts for assessing the relevance of each document for building a comprehensive event collection (i.e., recall-oriented and biased to documents with detailed background information) on a binary scale. Annotators had to follow these guidelines:

- Read the Wikipedia page of the event, to refresh the memory on the topic;
- Decide whether the central topic of the article is related to the event (by describing the event itself or a well-known premise/consequence);
- If yes, mark the document as relevant, otherwise as non-relevant. When undecided, mark it as non-relevant.

In order to examine the complexity of the task and measure the agreement between the two annotators, we initially ask them to annotate 250 documents from different datasets and regarding different types of events. The task is very time-consuming because the annotators often need to read the entire article before deciding on a relevance label. We come back on this issue when examining the trade-off between number of training data and performance of our supervised system.

Nevertheless, we obtain a good agreement between the two annotators with an inter-annotator agreement measured in Cohen's kappa of 0.78. The annotators assess the remaining dataset following the same approach. This leads to a gold standard of approximately 3700 documents annotated with binary judgments (33% of them as relevant, as presented in Table 5).

6.4.2 Baselines

Each method defines a query representation and then ranks the results according to the cosine similarity between the vector representations of the query and the document.

Event-name Retrieve documents that mention all the query words (e.g., "orange" and "revolution") and rank the results by TF-IDF cosine similarity. This is a common approach for building event collections [26].

Wikipedia Build a language model using words from the Wikipedia article of the event (e.g., https://en.wikipedia.org/wiki/Orange_Revolution) and rank by TF-IDF cosine similarity the documents in the sub-corpus.

Contextual Build a language model using the context passages (i.e., sentences) from the articles in the collections where the event is mentioned, and rank documents by TF-IDF cosine similarity.

Since our pipeline adopts different query expansion models (see Phase 6), we also examine the quality of each of these models individually. First of all, we consider **Place**, **Ent-TFIDF** and **Ent-GloVe**. In addition to this, we test the usefulness of adopting event aspects (see Phase 4) over entity aspects for entity-query expansion; we experiment both with TF-IDF and word embedding vector representations.

Approaches that use entity aspects are **EntAspTFIDF** and **EntAsp-GloVe**, while systems adopting event aspects are **EvAsp-TFIDF** and **Our-light**.[25]

[25] Which corresponds to **EvAsp-GloVe**.

The system presented by [19] is not applicable to our setting as it relies at its heart on focused crawling and is suitable for hyperlinked texts, which are not present in our corpora. Nevertheless, as the second step of their pipeline (i.e, "Topic Relevance Estimation") adopts TF-IDF cosine similarity to further filter the results, we investigate in our evaluation whether semantic representation of concepts and entities (i.e., word embeddings) could improve over symbolic features, such as tokens, and therefore be also useful for the work of [19].

6.4.3 Results on document retrieval

For each event, the systems offer a ranking of documents. We initially discuss the overall quality of the adopted methods; next we examine in detail the output of a difficulty test, a cost–benefit analysis of Our-full and the event-based performance.

Overall performance As a first step, we evaluate the quality of the ranking using trec_eval[26] and measuring the mean average precision (MAP) both on the New York Times Corpus and on the US Congressional Record Corpus. In Fig. 2[27] and 3 it is shown how the adoption of a simple document filtering approach such as retrieving the documents that mention the name of the event (**Event-name**) leads to poor results when compared to more advanced approaches. Additionally, we can notice how entity-query expansion approaches lead to good results, especially when representing the query as an embedding vector. This is an important finding, also relevant for the recent work of [19], which second step could benefit from the use of word embeddings as an alternative to TF-IDF feature vectors. Another important finding is that expanding the query in a coarse-grained way, using textual information directly extracted from Wikipedia or from initially retrieved document, leads to very poor performance, in comparison with more fine-grained query expansion approaches which use relevant entity and event aspects.

When retrieving relevant documents simply by using the location (i.e., **Place**), the results strongly differ between the two datasets. To better understand these results, consider the event Orange Revolution. As every day the New York Times publishes articles on global news, not all of the articles mentioning "Ukraine" will discuss the event, but they can also be about international deals or sport competitions. On the other hand, the US Congress mainly discusses issues regarding the United States internal and foreign affairs. Therefore, "Ukraine" will be mentioned only in a few particular cases, such as the outbreak of a large-scale protest.

Finally, a few takeaways regarding our system, which combines the outputs of different retrieval models with learning-to-rank. Firstly, in both collections the learning-to-rank method (**Our-full**) achieves the best results and, especially on the NYT Corpus, with a statistically significant improvement (paired t-test, significance level 0.01) over all other approaches. A second important outcome of the evaluation is that **Our-light** approach, when applied to the NYT Corpus, obtains statistically significant improvement over all the baselines. All methods (except "Place," as described above) show lower performance on the USC Corpus than on the NYT Corpus. This is because NYT articles are always about a specific topic, while this is not the case with USC speeches. Congressional speeches often address multiple topics and mention relevant entities out of context, such as part of comparisons, lists, or briefings.

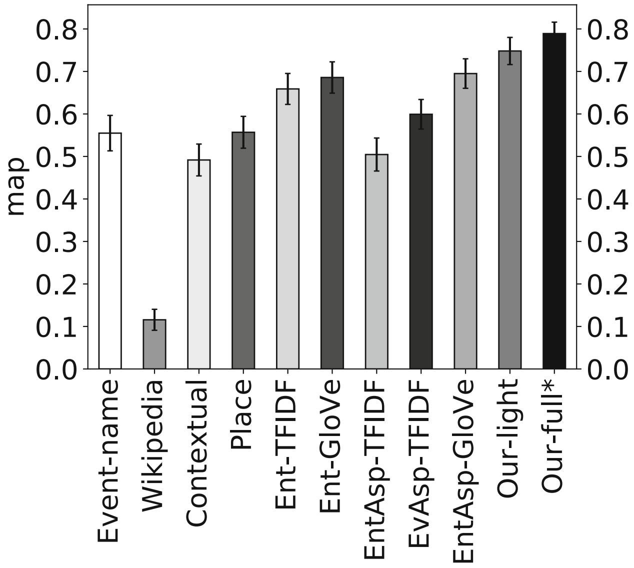

Fig. 2 MAP results on NYT Corpus

Corpus-based difficulty test After having measured the overall performance of the different methods, we examine the improvement of our approaches over a common heuristic for building event collections, namely using the Event-name. In order to do so, we present in Figs. 4 and 5 a comparison showing for each method the mean performance for queries of different difficulties. We divide the queries into different quartiles based on whether **Event-name** obtained good results (easy) or not (difficult), to analyze the different strengths and weaknesses of the methods.

If we consider the results on the New York Times Corpus, we can see that **Our-full** method performs better on all but the 5% easiest queries. Results on the US Congress show the complexity of building event collections on this corpus. However, we also see that our learning-to-rank approach is often able to benefit from the features used.

Costs and benefits of learning to rank As described above, the use of supervision in our learning-to-rank set-

[26] http://trec.nist.gov/trec_eval/.

[27] Method marked with * is significantly better than all others on its left.

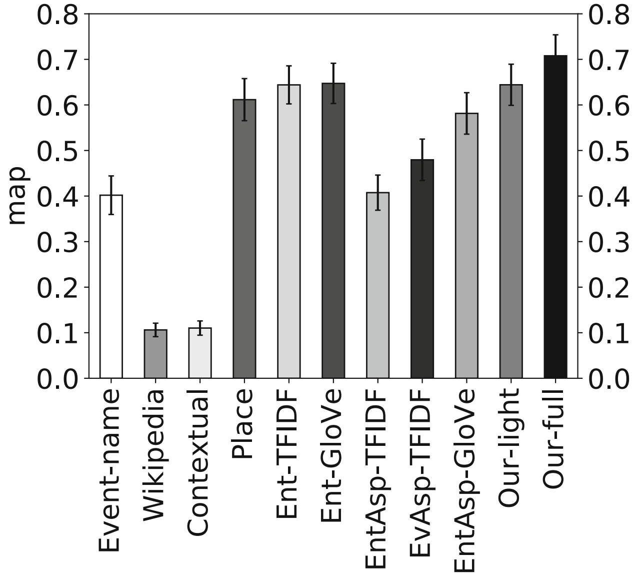

Fig. 3 MAP results on USC Corpus

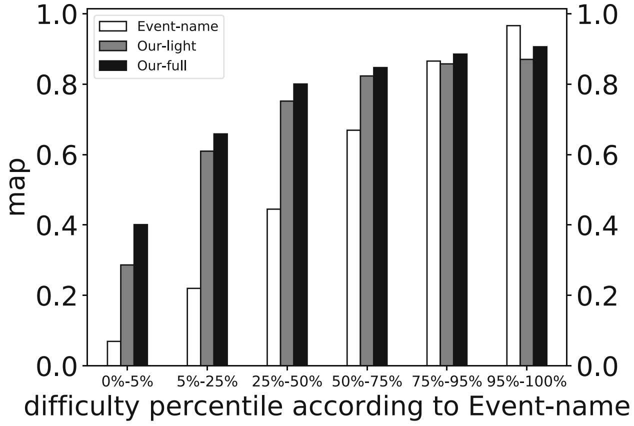

Fig. 4 Difficulty test on NYT Corpus

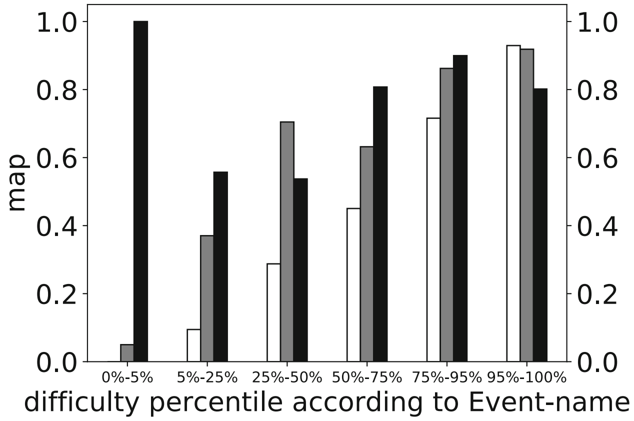

Fig. 5 Difficulty test on USC Corpus; column order as in Fig. 4

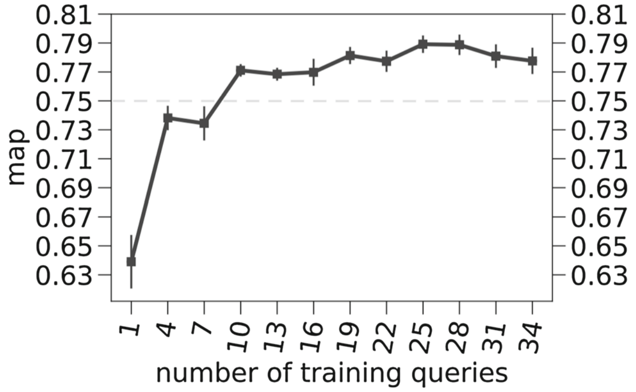

Fig. 6 Relation between MAP and number of training queries on NYT Corpus. Dashed line marks the performance of the best unsupervised baseline (Our-light)

ting (see Phase 7) boosts the performance of our system on both datasets. However, as we have discussed in subsection 6.4, this approach needs training data annotated by domain experts, which are expensive and time-consuming to generate. For instance, our NYT Gold Standard is composed by 44 event queries, each one associated with around 40 annotated documents, for a total of 1836 labeled articles (see statistics in Table 5). For establishing the relevance of each document, the annotators often needed to read the entire text, and therefore, this took more than 30 min of work per query, for a total of around 26 h of work per dataset.

Given this issue, in this subsection we study the trade-off between the amount of data used to train Our-full approach and its performance (in terms of MAP). We experiment with different numbers of queries as training data, varying between 1 and 34—i.e., the maximum number of training data used by our system in the fivefold cross-validation presented above. We randomly assign queries to the training set, and then we use all the remaining for testing. We repeat each test 50 times.

As shown in Fig. 6, already with 10 training queries (which means around 400 documents annotated in 5–6 h of work) Our-full system is able to outperform Our-light, namely the best unsupervised approach presented in Fig. 2. Additional queries permit to consistently reach a MAP of over 0.77.

We obtain similar findings on the USC Corpus (see Fig. 7), where already 7 annotated queries permit to outperform all the baselines.

To conclude, while it is time-consuming to generate these training data, already with 10 queries and 400 labeled documents our system is able to offer strong performance.

Event-based performance Given the findings presented above, as a final step of the evaluation we present a comparison between the baseline Event-name, the use of related entities and our method in its *light* and *full* versions, considering the three different types of events we employed as

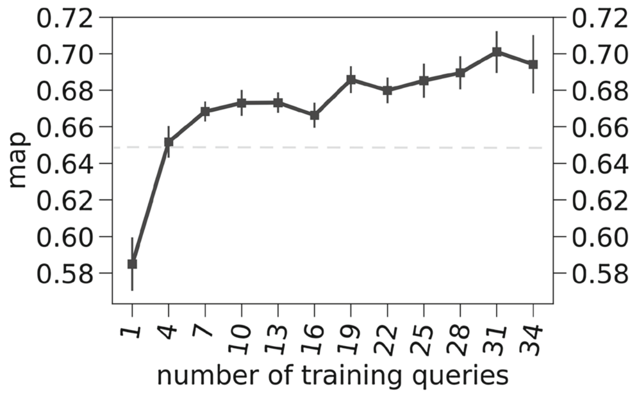

Fig. 7 Relation between MAP and number of training queries on USC Corpus. Dashed line marks the performance of the best unsupervised baseline (Our-light)

Table 6 MAP for different event types on the NYT Corpus

Method	Elections	Crises	Wars
Event-name	0.64 ± 0.06	0.39 ± 0.06	0.61 ± 0.06
Ent-TFIDF	0.63 ± 0.05	0.59 ± 0.06	0.76 ± 0.04
Our-light	0.72 ± 0.05	0.73 ± 0.06	0.83 ± 0.04
Our-full	**0.76 ± 0.04**	**0.74 ± 0.06**	**0.86 ± 0.04**

Bold indicates the higher number in each column

Table 7 MAP for different event types on the USC Corpus

Method	Elections	Crises	Wars
Event-name	0.32 ± 0.07	0.38 ± 0.06	0.52 ± 0.06
Ent-TFIDF	0.65 ± 0.07	0.63 ± 0.06	0.65 ± 0.06
Our-light	0.52 ± 0.06	**0.70 ± 0.05**	0.73 ± 0.06
Our-full	**0.73 ± 0.05**	0.63 ± 0.09	**0.77 ± 0.08**

Bold indicates the higher number in each column

queries in our work (unexpected elections, political crises and civil wars).

In Table 6, we report the results on NYT Corpus. Firstly, we can see how **Our-full** system always drastically improves over the **Event-name** baseline. In particular for political crises, we can see how the Event-name performance is more than 30% below the ones of Our-full system; this is due to the fact that the premises of a protest are complex to track, as a common name for the event is not yet established. (We expand on this in the next section.) Secondly, we notice that our approach achieves the best performance across all three event types. Finally, we remark that **Our-light** version of the system often provides as good rankings.

The results over the more complex USC Corpus (Table 7) show that our method always strongly improves over the Event-name baseline (at least 25% better on each type of event). In addition, we see how both elections and political crises are difficult to track, especially because both are often mentioned out of context.

Table 8 Percentage of documents missed using the Event-name heuristics on NYT Corpus

Type of event	Before	After
Elections	16% ± 6	22% ± 7
Crises	63% ± 9	31% ± 6
Wars	14% ± 4	8% ± 2
All	30% ± 5	20% ± 4

7 Discussion: temporal and large-scale

We present here a few findings regarding the advantages of using the system introduced in this paper over the commonly adopted Event-name baseline; finally, we examine its potential and drawbacks testing it on two different TREC resources.

Documents missed by Event-name heuristic At the heart of our approach lies the hunch that using the Event-name as a filtering method for building event collections has low performance in that this method is to be able to capture information about premises and background stories. We examine this issue experimentally on the NYT Corpus by considering the three types of event previously presented. The findings of this analysis are presented in Table 8.

First of all, it is important to remark that using **Event-name** leads to an overall loss of around 25% of the relevant documents. However, by evaluating the performance of this heuristic on documents from before the event, we see that on average 30% of documents are missed and, in the case of political crises, this is increased to a miss-rate of over 60%.

Fine-grained diachronic comparison In Fig. 8, we compare performance (MAP) across different time intervals (from 4 weeks before, to 4 weeks after), between the Event-name baseline and **Our-light** version of the system. We consider both the results obtained over all events and specifically regarding political crises. From Fig. 8, it is evident that the performance of the **Event-name** baseline is always lower than our system, especially for what concerns the premises and the early stages of the event. This is especially evident when considering only political crises, where the Event-name does not retrieve almost any relevant document in the weeks leading up to the event.

Performance on TREC KBA stream corpus As an additional study, we present a detailed error analysis of our system in a series of complex realistic scenarios on a very large corpus. We use the previously introduced TREC KBA Stream Corpus, one of the few large-scale web archives fully available for research. It is composed of news and social media posts and spans for 15 months (October 2011–February 2013).

We consider five protests/crises that happened in this period: the Port Said Stadium Riot, the In Amenas Hostage

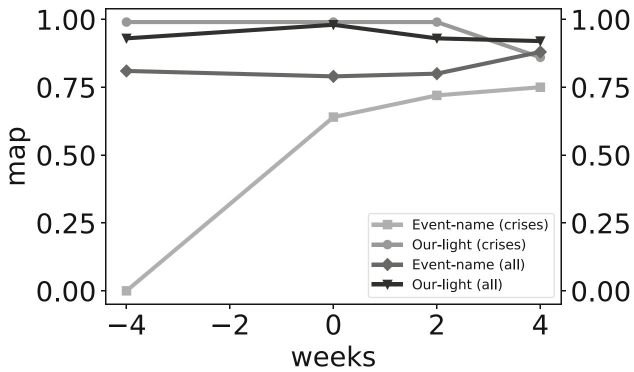

Fig. 8 MAP per time intervals comparing the performance of Event-name and Our-light on the NYT Corpus, regarding all events and only political crises

Table 9 Average precision on KBA Corpus

Event	Ev-name	Our-light	Our-full
Port Said St. riot	0.00	0.33	**0.92**
In Amenas crisis	0.00	**0.33**	0.13
Shahbag protest	0.00	**1.00**	0.85
Occupy Nigeria	**0.73**	0.44	0.68
Idle No More	0.00	0.16	**0.52**
MAP	0.14	0.45	**0.62**

Bold indicates the higher number in each column

Crisis, the 2013 Shahbag Protests, Occupy Nigeria and Idle No More. We examine the performance of our system for retrieving documents on the premises and the early stages of these events (i.e., from 4 weeks before, until the day of the event). After having assessed the overall quality of the ranking and the improvement over the Event-name (see Table 9),[28] we have conducted an in-depth error analysis.

The quality of the output of our system varies a lot across the events. For two of them, it leads to very good results, retrieving all relevant documents on high ranks. These are events characterized by a precise location (the Port Said riots in the stadium) or that received large coverage in international news (the Shahbag protests in Bangladesh). However, crises that overlap with other events happening at the same time in the same place (e.g., the In Amenas Hostage Crisis during the discussions on closing the border between Algeria and Mali) are much more difficult to track. This evaluation also reconfirms that the **Event-name** is a good retrieval approach only when the protest has a name from its early stages onward, as for Occupy Nigeria.

An extreme example of the difficulties of the task concerns the retrieval of documents regarding small-scale grassroots movements, such as the Canadian protest Idle No More, in a corpus of international news. This event, in its early stages, has only few relevant documents in the corpus. While our system retrieves these relevant documents at the top positions of the ranking, not a single relevant document is retrieved using the Event-name baseline. This is because the phrase "Idle No More" is not mentioned within these documents.
These final experiments demonstrate that the advantages of our system over the Event-name baseline translate to a large-scale corpus of multiple terabyte.

Supporting Wikipedia enrichment As a last experiment before concluding this work, we examine the usefulness of our approach for supporting the enrichment of event articles on Wikipedia, by retrieving additional information. To test our system, we employ the new TREC Complex Answer Retrieval (CAR) Dataset [14], on which one of the sub-tasks presented by the organizers focuses on assigning a textual paragraph to the related Wikipedia article. We created a collection of 1285 paragraphs, extracted from all event pages that we studied in our work. Then, we compare the performance of the Event-name heuristic against the use of entities and entity aspects, in order to associate each paragraph with the correct event page. Given the fact that Our-full approach relies on the use of event aspects (i.e., the same paragraphs that we aim to retrieve from the collection), we cannot employ it in this task.

Nevertheless, as can be seen in Table 10, the use of **entities** and **entity aspects** is a strong alternative to Our-full approach and shows a significant improvement over the Event-name heuristics. In addition to this, the results obtained are in line with findings previously presented on the Trec CAR dataset [39], especially for what concerns the usefulness of supporting passages for query expansion. It is also important to note that, as opposed to the results obtained on NYT and USC corpora, when dealing with Wikipedia content, TF-IDF vector representations perform better than word embeddings. The same finding also emerged in previous work on TREC-CAR [39] and could be related to the presence of strong lexical similarities across Wikipedia pages, which are easily captured through the use of term-frequency analyses.

We therefore conclude that the system presented in this paper could also be useful for supporting the enrichment of event articles on Wikipedia by retrieving information on its related aspects, such as premises or consequences.

8 Conclusion

In this journal extension of our previous work [41], we expanded the presentation of a system for creating event collections from large datasets. This approach selects not only the core documents related to the event itself, but most impor-

[28] We detected and removed news duplicates from the initial pool of potentially relevant documents, before conducting the final evaluation.

Table 10 Mean Average Precision, Reciprocal Precision and Mean Reciprocal Rank of the approaches on TREC CAR

Method	MAP	R-Prec	MRR
Random baseline	0.03	0.02	0.07
Event-name	0.40	0.39	0.81
Ent-TFIDF	0.59	**0.59**	**0.93**
Ent-GloVe	0.47	0.46	0.83
EntAsp-TFIDF	**0.63**	**0.59**	0.80
EntAsp-GloVe	0.50	0.44	0.72

tantly includes documents which describe related aspects, such as premises and consequences. We do so through the use of relevant entities and textual passages (i.e., event aspects), which are collected from a knowledge base, and whose similarity to the documents examined is interpreted as one of many indicators of relevance.

We evaluate our system on different diachronic collections studying various types of events, such as unexpected elections, political crises and civil wars. In particular, we show how in all contexts, our approach consistently improves over the use of the Event-name heuristic for building event collections. We evaluate different methods including the use of word embeddings and TF-IDF, information from entity's articles and passages surrounding entity links.

The best single method uses embedding representations of relevant entities and event aspects to expand the query. This approach, depending on the collection and event type, is able in some cases to already obtain good performance. Using this method together with several variants in a learning-to-rank framework brings additional improvements in the remaining cases. We provide evidence that our method is capable of identifying documents from the early stages of an event, when the name is not yet established. We test our approach extensively on the New York Times and US Congressional Record corpora and demonstrate that our results generalize to other collections such as the TREC-CAR dataset and the TREC-KBA Stream corpus.

Given its potential for creating comprehensive event collections, our system can now sustain humanities and social science researchers when dealing with the vastness of born-digital materials.

Acknowledgements This work was funded by a scholarship of the Eliteprogramm for Postdocs of the Baden-Württemberg Stiftung (project "Knowledge Consolidation and Organization for Query-specific Wikipedia Construction") and by an AWS Research Award (Promotional credits name: EDU_R_FY2015_Q3_MannheimUniversity_Dietz). Furthermore, this work was also supported by the Junior-professor funding program of the Ministry of Science, Research and the Arts of the state of Baden-Württemberg (project "Deep semantic models for high-end NLP application").

References

1. Abujabal, A., Berberich, K.: Important events in the past, present, and future. In: WWW (2015)
2. Allan, J.: Introduction to topic detection and tracking. In: Allan, J. (ed.) Topic Detection and Tracking. The Information Retrieval Series, vol. 12. Springer, Boston, MA (2002)
3. Allan, J., Lavrenko, V., Jin, H: First story detection in TDT is hard. In: CIKM (2000)
4. Allan, J., Papka, R., Lavrenko, V.: On-line new event detection and tracking. In: SIGIR (1998)
5. Aslam, J.A., Ekstrand-Abueg, M., Pavlu, V., Diaz, F., Sakai, T.: TREC 2013 temporal summarization. In: TREC (2013)
6. Au Yeung, C.-M., Jatowt, A.: Studying how the past is remembered: towards computational history through large scale text mining. In: CIKM (2011)
7. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A Nucleus for a Web of Open Data. Springer, Berlin (2007)
8. Bailey, S., Thompson, D.: Building the uk's first public web archive. D-Lib **12**, 1 (2006)
9. Bethard, S.: Cleartk-timeml: a minimalist approach to tempeval 2013. In: SEM (2013)
10. Blanco, R., Ottaviano, G., Meij, E.: Fast and space-efficient entity linking for queries. In: WSDM (2015)
11. Cano, I., Singh, S., Guestrin, C.: Distributed non-parametric representations for vital filtering: UW at TREC KBA. In: TREC (2014)
12. Ceroni, A., Gadiraju, U., Matschke, J., Wingert, S., Fisichella, M.: Where the event lies: predicting event occurrence in textual documents. In: SIGIR (2016)
13. Dalton, J., Dietz, L., Allan, J.: Entity query feature expansion using knowledge base links. In: SIGIR (2014)
14. Dietz, L., Gamari, B.: TREC CAR: A Data Set for Complex Answer Retrieval. Version 1.5 (2017)
15. Eisenstein, J., O'Connor, B., Smith, N.A., Xing, E.P.: A latent variable model for geographic lexical variation. In: EMNLP (2010)
16. Ferragina, P., Scaiella, U.: Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In: CIKM (2010)
17. Glavaš, G., Šnajder, J.: Construction and evaluation of event graphs. Nat. Lang. Eng. **21**, 04 (2015)
18. Gomes, D., Miranda, J., Costa, M.: A survey on web archiving initiatives. In: TPDL (2011)
19. Gossen, G., Demidova, E., Risse, T.: Extracting event-centric document collections from large-scale web archives. In: TPDL (2017)
20. Graus, D., Peetz, M.-H., Odijk, D., de Rooij, O., de Rijke, M.: yourhistory-semantic linking for a personalized timeline of historic events. In: Workshop: LinkedUp Challenge at OKCon (2013)
21. Gupta, D.: Event search and analytics: detecting events in semantically annotated corpora for search and analytics. In: WSDM (2016)
22. Hasibi, F., Balog, K., Bratsberg, S.E.: Exploiting entity linking in queries for entity retrieval. In: ICTIR (2016)
23. Hockx-Yu, H.: Access and scholarly use of web archives. Alex. J. Nat. Int. Libr. Inf. **25**, 1–2 (2014)
24. Hyde, S.D., Marinov, N.: Which elections can be lost? Polit. Anal. **20**, 191–210 (2012)
25. Jatowt, A. Au Yeung, C.-M.: Extracting collective expectations about the future from large text collections. In: CIKM (2011)
26. Kedzie, C., McKeown, K., Diaz, F.: Summarizing disasters over time. In: Workshop on Social Good at SIGKDD (2014)
27. Kotov, A., Zhai, C.: Tapping into knowledge base for concept feedback: leveraging conceptnet to improve search results for difficult queries. In: WSDM (2012)
28. Kuzey, E., Vreeken, J., Weikum, G.: A fresh look on knowledge bases: distilling named events from news. In: CIKM (2014)

29. Lepore, J.: The cobweb: can the internet be archived? The New Yorker (2015)
30. Lewis, D.: The trec-4 filtering track. In: TREC (1995)
31. Li, H.: Learning to rank for information retrieval and natural language processing. Synth. Lect. Hum. Lang. Technol. **7**, 3 (2014)
32. Liu, X., Fang, H.: Latent entity space: a novel retrieval approach for entity-bearing queries. Inf. Retr. J. **18**, 6 (2015)
33. Lyman, P., Kahle, B.: Archiving digital cultural artifacts. D-Lib **4**, 7 (1998)
34. Menini, S., Sprugnoli, R., Moretti, G., Bignotti, E., Tonelli, S., Lepri, B.: Ramble on: tracing movements of popular historical figures. In: EACL (2017)
35. Miller, G.A.: Wordnet: a lexical database for english. Commun. ACM **38**, 11 (1995)
36. Milligan, I., Ruest, N., Lin, J.: Content selection and curation for web archiving: the gatekeepers vs. the masses. In: JCDL (2016)
37. Mishra, A., Berberich, K.: Expose: exploring past news for seminal events. In: WWW (2015)
38. Mishra, A., Berberich, K.: Event digest: a holistic view on past events. In: SIGIR (2016)
39. Nanni, F., Mitra, B., Magnusson, M., Dietz, L.: Benchmark for complex answer retrieval. In: ICTIR (2017a)
40. Nanni, F., Ponzetto, S.P., Dietz, L.: Entity relatedness for retrospective analyses of global events. In: NLP+CSS at WebSci (2016)
41. Nanni, F., Ponzetto, S.P., Dietz, L.: Building entity-centric event collections. In: JCDL (2017b)
42. Nanni, F., Ponzetto, S.P., Dietz, L.: Entity-aspect linking: providing fine-grained semantics of entities in context. In: JCDL (2018)
43. Nanni, F., Zhao, Y., Ponzetto, S.P., Dietz, L.: Enhancing domain-specific entity linking in dh. In: DH (2017c)
44. Ntoulas, A., Cho, J., Olston, C.: What's new on the web? In: WWW (2004)
45. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: EMNLP (2014)
46. Pound, J., Mika, P., Zaragoza, H.: Ad-hoc object retrieval in the web of data. In: WWW (2010)
47. Raviv, H., Kurland, O., Carmel, D.: Document retrieval using entity-based language models. In: SIGIR (2016)
48. Ristoski, P., Paulheim, H.: Rdf2vec: Rdf graph embeddings for data mining. In: ISWC (2016)
49. Rollason-Cass, S., Reed, S.: Living movements, living archives: selecting and archiving web content during times of social unrest. N. Rev. Inf. Netw **20**, 1–2 (2015)
50. Rovera, M., Nanni, F., Ponzetto, S.P., Goy, A.: Domain-specific named entity disambiguation in historical memoirs. In: CLiC-it (2017)
51. Schich, M., Song, C., Ahn, Y.-Y., Mirsky, A., Martino, M., Barabási, A.-L., Helbing, D.: A network framework of cultural history. Science **345**, 6196 (2014)
52. Schuhmacher, M., Dietz, L., Paolo Ponzetto, S.: Ranking entities for web queries through text and knowledge. In: CIKM (2015)
53. Singh, J., Nejdl, W., Anand, A.: Expedition: a time-aware exploratory search system designed for scholars. In: SIGIR (2016)
54. Sprugnoli, R., Tonelli, S.: One, no one and one hundred thousand events: defining and processing events in an inter-disciplinary perspective. Nat. Lang. Eng. **23**, 485 (2016)
55. Tuck, J.: Web archiving in the UK: cooperation, legislation and regulation. Liber Q. **18**, 3–4 (2008)
56. Witten, I., Milne, D.: An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In: Workshop on Wikipedia and Artificial Intelligence at AAAI (2008)
57. Wolfreys, J.: Readings: Acts of Close Reading in Literary Theory. Edinburgh University Press, Edinburgh (2000)
58. Xiong, C., Callan, J.: Esdrank: connecting query and documents through external semi-structured data. In: CIKM (2015)