Building Entity-Centric Event Collections

Federico Nanni, Simone Paolo Ponzetto
Data and Web Science Group
University of Mannheim
Germany
federico,simone@informatik.uni-mannheim.de

Laura Dietz
Department of Computer Science
University of New Hampshire
USA
dietz@cs.unh.edu

ABSTRACT

Web archives 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 1) identifying relevant concepts and entities from a knowledge base, and 2) 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, and we test its performance on the TREC KBA Stream corpus, a large and publicly available web archive.

ACM Reference format:
Federico Nanni, Simone Paolo Ponzetto and Laura Dietz. 2016. Building Entity-Centric Event Collections. In *Proceedings of ACM Conference, Toronto, Canada, June 2017 (JCDL'17),* 10 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

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 Economic Crisis.

Given the known ephemerality of born-digital materials [26, 36], since the 90s, public and private institutions have embraced the responsibility of preserving these resources for future studies [16]. While web archives such as the Internet Archive [30] 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 [20] – it is infeasible for scholars to conduct close reading analyses [47] of specific topics. 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 this work we focus on the task of automatic event-collection building from large corpora of previously harvested documents (such as news, transcript of political speeches 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 analysed by a historian, for example, 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.

Our contribution. In order to achieve this goal, we propose an 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-depth evaluation. We evaluate the presented system on three different datasets, using several well-known reference baselines and separate evaluations of different components for entity, passage, and document selection. We provide an in-depth analysis with respect to entity selection, passage analysis, and document filtering 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.[1] We further include *c)* a third dataset, a large (10TB) and publicly available web archive, namely the TREC KBA Stream corpus, which includes both news and social media posts, collected between 2011 and 2013.

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

[1] All gold standards available at: https://federiconanni.com/event-collections/

978-1-5386-3861-3/17/$31.00 ©2017 IEEE

2 BACKGROUND: EVENT-COLLECTION BUILDING

The task of building event collections from large corpora, which we tackle in this paper, 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 [32].[2] 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 [41, 45]. These collections are created and curated by "the Archive-It team in conjunction with curators and subject matter experts from institutions around the world".[3] The same approach has been employed by public institutions.[4]

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. 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 missing publishing selection guidelines (what to include and what not).

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 summarisation task organised 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 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 free Algerian elections since independence (1990), which is 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.

[2] See for example Nick Ruest collection of the Bataclan attack: http://ruebot.net/post/look-14939154-paris-bataclan-parisattacks-porteouverte-tweets
[3] More info here: https://archive-it.org/organizations/89
[4] For example, the UK Web Archive: https://www.webarchive.org.uk/ukwa/collection

3 RELATED WORK

The task we address in this paper 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 twenty 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 [27]. In this task, a stream of documents is to be filtered to documents about a given information need. More recently, the TREC Knowledge Base Acceleration track began to study how to track people and organisations in a diachronic collection by building language models of entities that change over time [10].

Early efforts on tracking events in a stream of news were made in the Topic Detection and Tracking Task (TDT) at the Text Retrieval Conference [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 Summarisation track has aimed to provide introspective passage summarisations of an event as it is unfolding [23].

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 "Mr Miller went to Boston in August" connects the entities Mr Miller, Boston and August with the predicate "went to". During the last decade, thanks to the efforts in developing annotation guidelines, conducting evaluation campaigns[5] and organising specific workshops[6], 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 [8, 15].

Given the importance of events as a topic of study in historical research, Sprugnoli and Tonelli [44] 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 recognising '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

[5] https://www.ldc.upenn.edu/collaborations/past-projects/ace; http://www.timeml.org/tempeval/
[6] For example: https://sites.google.com/site/cfpwsevents/

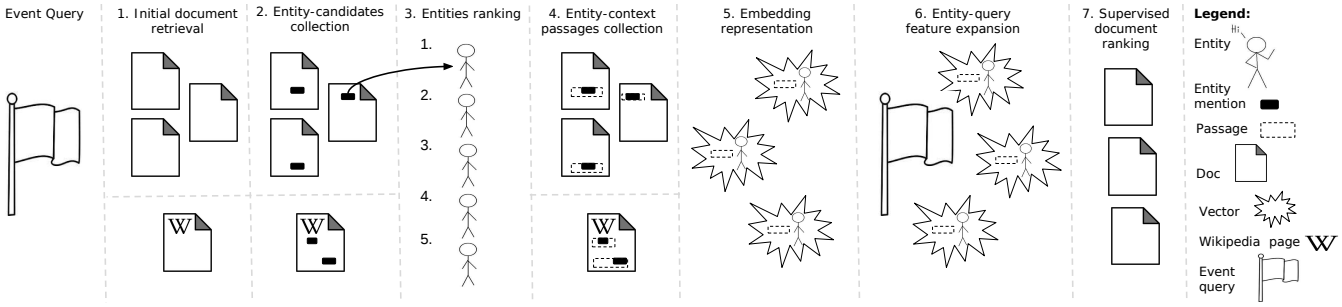

Figure 1: Pipeline schema of our system.

event-collections are created by public and private archival institutions, they are also generally built around named-events (e.g., the Charlie Hebdo Shooting[7]). For these reasons, in our work we focus on building collections for a given named-event.

Events and entities. The importance of employing geographical [13] and temporal [22] 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 [17], generating event-digests from Wikipedia [34] and building time-aware exploratory search systems [11, 43] (often considering the name of the event as the query [33]). The task of extracting important events adopting named entities has been recently addressed by Abujabal and K. Berberich [1] and by Gupta [18]. Entities have also been used to study the general perception of society towards past events [6]. Kuzey et al. [25] employ named entities to extract yet unknown named events for knowledge 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 [24, 31], 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 [9, 19, 39]. Combinations of knowledge base retrieval and entity linking methods have been studied for web search queries both for entity ranking tasks [38, 42] as well as document ranking tasks [12, 29, 48]. 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 Figure 1. The user selects a named-event V of interest, such as the 2004 Orange Revolution and a specific collection C, for example the New York Times Corpus. As remarked above, the named-event is expected to be an entity in DBpedia [7].

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

Phase 1: Initial document retrieval: Retrieve an initial set of documents D from C using the name of the event as a query Q and collect all the documents with a mention M of the event name.

Phase 2: Entity-candidates collection: Extract a set of potentially relevant entities E from two resources: the pool of relevant documents D and the Wikipedia page W of the event. Entities from D are extracted using the entity linker TagMe2 [14] and collecting the entities in the surrounding context of the event mentions M (i.e., in a window of three sentences). Entities from W are collected following all Wikipedia outlinks. This approach is inspired by work on entity query feature expansion [12, 29].

Phase 3: Entities ranking. Rank entities E by relevance to the event, leveraging information from the knowledge base. As a rank measure for entity-event relatedness, we use the cosine similarity of vector representation for each entity and event taken from the RDF graph embedding representations provided by Ristoski et al [40] – using a 500-dimensional vector space.

Phase 4: Entity-context passages collection. For each relevant entity E, collect a text passage P from the Wikipedia page W of the event by retrieving the first passage (i.e., three sentences) that contains a link to E. In case E does not appear in W, retrieve P from the collected relevant documents D.

Phase 5: Embedding representation. Project entities E and contextual passages P into the embedding space, in order to obtain their latent vector representations (GE and GP). Do this by computing the element-wise average of the embeddings of the E and P. Let S be for example the set of unique words in P. The embedding of P (GP) is then computed as follows:

$$GP = \frac{1}{N} \sum_{w \in S} \text{freq}(w) \cdot \vec{v}_w$$

where $\text{freq}(w)$ is the frequency with which word w occurs in P, v_w is the embedding vector of the word w, and N is the total number of words in P. The same is applied to obtain GE. We use the state-of-the-art pre-computed GloVe word embeddings (300d) [37].

Phase 6: Entity-query feature expansion. Expand the initial event query Q with the following vector-space expansion models. We interpret entities as words, e.g. entity "Yulia Tymoshenko" is represented as the words "Yulia" and "Tymoshenko". We represent each word as a vector, and build an expanded query vector representation from the element-wise sum of these vectors. The results are ranked according to the cosine similarity of query and

document vector. We study two variations of vector space models: TF-IDF (logarithmic, L2-normalised variant) over the corpus vocabulary and the GloVe word embedding (embed).

- **Place.** Expansion with only the location entity L (e.g., "Kiev") using a TF-IDF vector. We argue that, in specific cases, the location is already a precise indicator for retrieving relevant documents.
- **Entities.** Expansion with top 10 related entities E as ranked in Phase 3, using TF-IDF.
- **Ent+Pass.** Expansion with words from contextual passages P as collected Phase 4 of top 10 related entities, using TF-IDF vector representation.
- **Emb-Ent.** Expansion with top 10 related entities E as ranked in Phase 3, using GloVe vector representation.
- **Emb-Ent+Pass/ Our light.** Expansion with words from contextual passages P as collected in Phase 4 of top 10 related entities, using GloVe vector representation. This is our championed method which we refer to as **our light** in the remainder of this work.

Phase 7: Supervised document ranking / Our full. Combine the ranking-score of different methods studied in Phase 6 with supervised machine learning in a learning-to-rank setting [28], for producing a final ranking of relevant documents. For evaluation we perform training/testing with 5-fold cross validation; we refer to it as **our-full**.

5 EXPERIMENTAL SETUP

In this section we introduce the collection of sources and case studies where we evaluate our approach for building event collections.

5.1 Datasets

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

NYT corpus. The New York Times Corpus comprises over 1.8 million articles published between 1987 and 2007.[8]

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.[9] The obtained 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 the Extensions of remarks. This collection sums up to over 1.2 million documents.

KBA corpus. For a final large-scale experiment, in the discussion part we consider a third dataset, namely the 2014 TREC KBA Stream Corpus, which is 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.

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), 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 summarisation), 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[10] as well as the result of a political crisis.[11] We identified 15 unexpected elections, which happened between 1989 and 2007 using the National Elections Across Democracy and Autocracy (NELDA) dataset[12] [21] and in particular by considering elections flagged with the variables NELDA 2 ("Were these the first multiparty 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?").[13]

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.[14]

Civil wars. The third type is civil wars. While tracking events such as wars between different countries 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 and often arise as a consequence of previous long-term political tensions inside the country. Therefore, we argue, these tensions can not be easily captured by simply searching for documents that mention the name of the event (e.g. Bosnian War).

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

[9] 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, 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).

[10] See for example the first multiparty election in Algeria, 1991.

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

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

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

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

We identify 14 civil wars, combining information from the NELDA dataset with a set of Wikipedia categories on the topic.[15]

6 EXPERIMENTAL EVALUATION

In this section, we first evaluate the quality of the approach we adopt for extracting and ranking entities that are related to an event. Next, we establish the quality of the extracted 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 and Ranking Entities

As our approach distinguishes collecting (see Phase 2) and ranking (see Phase 3) entities, we study the performance of each component in isolation. Given a named event, such as an election, an internal conflict or an anti-establishment protest, we compare our method 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 sub-set 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 by 830 annotated entity-event pairs (484 relevant and 346 not relevant) extend the gold standard of entity-event relatedness assessments we created for a previous work [35].

6.1.2 Compared Methods for Collecting Entities. In Phase 2, our system retrieves an initial 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 an manual (upperbound) reference baseline.

6.1.3 Compared Methods for Ranking Entities. In Phase 3 we rank entities by computing the cosine similarity between the RDF embeddings representation of each entity and event; as in Ristoski et al. [40], we call this method **RDF2Vec**. We study its performance in comparison with the following baseline 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.[16] We report the results computed on the NYT Corpus.

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

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

Table 1: Precision, recall and F1-Score regarding 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**

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 normalisation) 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.
- Rank entities according to the **ContFreq** baseline.

This method was used in our previous work [35] and is inspired by the work of Milne and Witten [46].

6.1.4 Results on Entities Collection. For each event, the different approaches for collecting potentially relevant entities present a set of candidates. Given our gold standard annotations, in Table 1 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 analysing 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 were not related.

Take-away. To conclude, while using Wikipedia Outlinks leads to good results, the best performance are obtained when creating a pool of entities by combining candidates collected from Wikipedia and candidates retrieved from contextual passages, and therefore we use this approach for Phase 2 of our system. This finding is in line with experiments of Dalton et al. [12].

6.1.5 Results on Entities Ranking. We evaluate the quality of the rankings using the mean-average precision metric (MAP). Additionally we report the micro-averaged precision at different cutoffs (5 and 10). The results are presented in Table 2. 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 we previously presented [35] and RDF2Vec show that, while our low-cost approach yields to good results, RDF2Vec clearly outperforms it. For this reason, we use RDF2Vec for Phase 3 of our system.

Table 2: Mean Average Precision and P@k regarding entity ranking.

Method	MAP	P@5	P@10
ContFreq	0.22 ± 0.03	0.55 ± 0.06	0.48 ± 0.05
CheapEntRel	0.51 ± 0.05	0.70 ± 0.05	0.62 ± 0.05
RDF2Vec	**0.65 ± 0.05**	**0.80 ± 0.06**	**0.74 ± 0.05**

6.2 Collecting Contextual Passages

The next step of our work is to collect passages where each relevant entity E is presented in the context of the event V. We compare the approach we adopted (see Phase 4) to other baselines.

6.2.1 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 elaborates on 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 [35].

6.2.2 Compared Methods. As described in Phase 4, we retrieve passages with the entity in the context of the event from the Wikipedia page of the event. We call this method **Wiki-Pass**. We compare its performance 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 3).

6.2.3 Results. For each entity, the different approaches for collecting potentially relevant passages present a candidate. Using our gold standard annotations, we report in Table 3 the precision, recall and F1-Score of the different approaches.

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 emerge 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 from the Wikipedia page of the event (**Wiki-Pass**) remains therefore the most efficient approach for the task, and therefore we use this approach in Phase 4.

6.3 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 method.

Table 3: Precision, recall and F1-Score regarding 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
Wiki-Pass	**0.99 ± 0.03**	0.81 ± 0.03	**0.89 ± 0.03**

6.3.1 Gold Standard. For each event we consider an initial pool of documents in each corpus as a starting subcorpus. These documents have been selected following these two premises: *a)* they are published maximum 18 months before or after the event (i.e., in 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 subcorpus, 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, then retain the top 15 documents in each ranking for manual assessment. We ask two domain experts to assess 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. 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).

6.3.2 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 [23].

Wikipedia. Build a language model using words from the Wikipedia article of the event (e.g., /wiki/Orange_Revolution) and rank by TF-IDF cosine similarity the documents in the subcorpus.

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 document retrieval models (see Phase 6), we also examine the quality of each of these models individually, namely: **place**, **entities**, **ent+pass** and **emb-ent**.

6.3.3 Results. For each event, the different 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 and the event-based performance.

Overall performance. As a first step, we evaluate the quality of the ranking using trec_eval[17] and measuring the mean average precision (MAP) both on the New York Times Corpus and on the US Congressional Record Corpus. In Figures 2 and 3 it is shown how the adoption of a document filtering approach such as retrieving the documents that mention the **event-name** leads to poor results, when compared to almost all the other approaches. Additionally, we can notice how entity-query expansion approaches, with the exception of the entities+passage method, always lead to good results, especially when representing the query as an embedding vector. 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 to more fine-grained query expansion approaches which use relevant entities and passages.

For what concerns 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 take-aways 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[18] 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 baselines.

To conclude, it is important to remark on the fact that all methods (except "place", as described above) have shown 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.

Corpus-based difficulty test. After having measured the overall performance quality of the different methods, we examine the improvement of our approach over a common heuristic for building event collections, namely using the event-name. In order to do so, we present in Figures 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 analyse 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

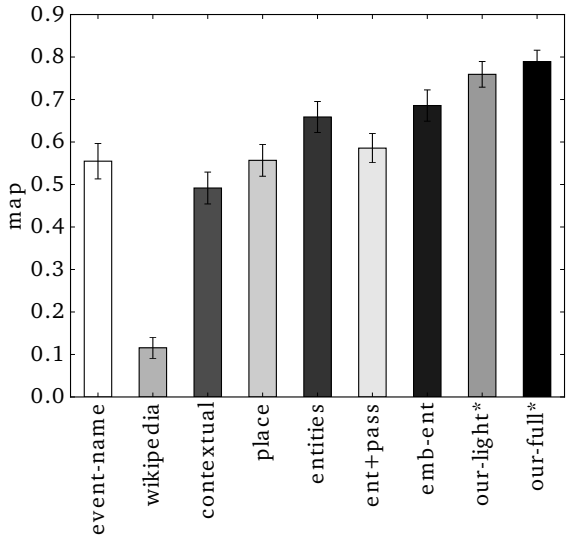

Figure 2: MAP results on NYT Corpus. Methods marked with * are significantly better than all others on their left.

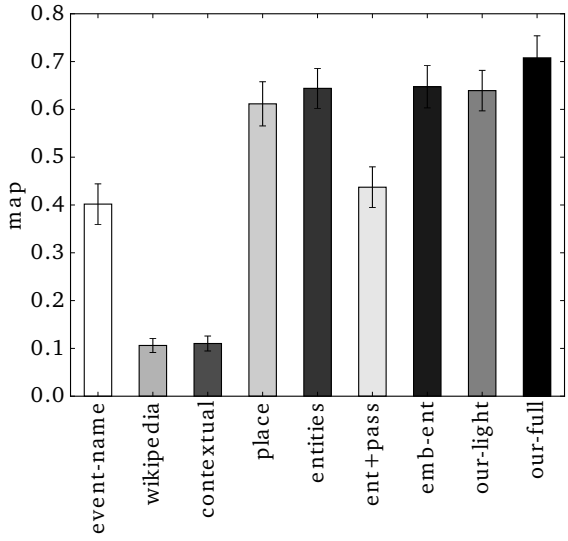

Figure 3: MAP results on USC Corpus.

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

[18] Paired t-test, significance level 0.01

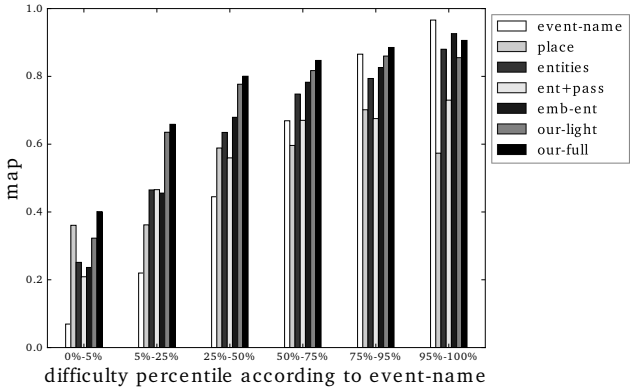

Figure 4: Difficulty Test on NYT Corpus.

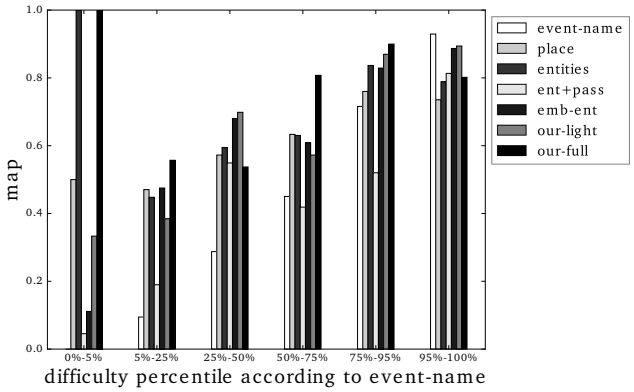

Figure 5: Difficulty Test on USC Corpus.

that different methods have different strengths on different query subsets and that our learning-to-rank final step (see our-full, Phase 7) is often able to benefit from it.

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 to expand the query and our method in its *full* and *light* version, considering the three different types of events we employed as queries in our work (unexpected elections, political crises and civil wars).

In Table 4 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 are over 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 Section 6). 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 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

Table 4: MAP for types of events on the NYT Corpus.

Method	Elections	Crises	Wars
event-name	0.64 ± 0.06	0.39 ± 0.06	0.61 ± 0.06
entities	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**

Table 5: MAP for types of events on the USC Corpus.

Method	Elections	Crises	Wars
event-name	0.32 ± 0.07	0.38 ± 0.06	0.52 ± 0.06
entities	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**

because both are often mentioned out of context. For example, a document about the political situation in Ethiopia says:

> The popular opposition to Ethiopia's current corrupt regime is comparable to the Orange Revolution in Ukraine and the brave Lebanese demonstrators who removed the Syrian puppet regime in their country.[19]

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 on a large-scale web archive.

Documents missed by event-name heuristic. The initial assumption on which our work has been based is that using the event-name as a filtering method for building event collections is not the ideal approach, due to the fact that information on premises and background stories could be missed. We examine this issue on the NYT Corpus by considering the three types of event previously presented. The findings of this analysis are presented in Table 6.

First of all, it is important to remark that using the **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 Figure 6, 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 Figure 6, it is evident that the performance of the **event-name** baseline are 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

[19] https://www.congress.gov/crec/2005/11/09/CREC-2005-11-09-pt1-PgE2308.pdf

Table 6: **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

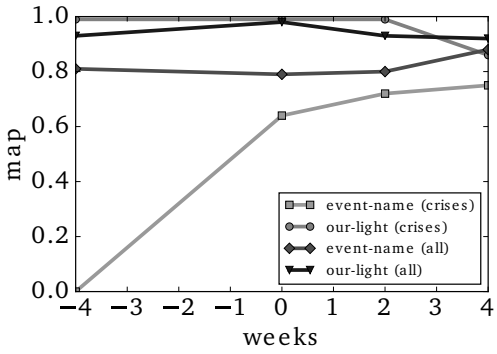

Figure 6: MAP per time intervals comparing the performance of event-name and our-light on the NYT Corpus, regarding all events and only political crises.

the event-name does not retrieve almost any relevant documents in the weeks leading up to the event.

Performance on TREC KBA stream corpus. We finalise our analysis with 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 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 four 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 7)[20], 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 characterised 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 which 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

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

Table 7: **Average Precision on KBA Corpus.**

Event	event-name	entities	our-light	our-full
Port-Said St. riot	0.00	**1.00**	0.33	0.92
In Amenas crisis	0.00	0.04	**0.33**	0.13
Shahbag protest	0.00	0.07	**1.00**	0.85
Occupy Nigeria	**0.73**	0.54	0.44	0.68
Idle No More	0.00	**1.00**	0.16	0.52
MAP	0.14	0.53	0.45	**0.62**

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 within 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 in the content of 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 tera byte.

8 CONCLUSION

In this paper we present a system for creating event collections from large datasets. Our approach selects not only the core documents related to the event itself, but most importantly includes documents which describe related aspects, such as premises and consequences. We do so through the use of relevant entities, which are collected from a knowledge base, and whose presence in text 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 contextual passages to expand the query. This approach, depending on the collection and event type, already obtains good performance in some cases. Using this methods together with several variants in a learning-to-rank framework brings additional improvements in the remaining cases. We provide evidence that our method is able to identify 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 generalise to large collections such as the TREC 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 collections.

Aknowledgements

This work was funded in part 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, with promotional credits name EDU_R_FY2015_Q3_MannheimUniversity_Dietz. Furthermore, this work was partially funded by the Juniorprofessor funding programme 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] Abdalghani Abujabal and Klaus Berberich. 2015. Important Events in the Past, Present, and Future. In *WWW*.
[2] James Allan. 2002. Introduction to topic detection and tracking. In *Topic detection and tracking*. Springer.
[3] James Allan, Victor Lavrenko, and Hubert Jin. 2000. First story detection in TDT is hard. In *CIKM*. ACM.
[4] James Allan, Ron Papka, and Victor Lavrenko. 1998. On-line new event detection and tracking. In *SIGIR*.
[5] Javed A Aslam, Matthew Ekstrand-Abueg, Virgil Pavlu, Fernando Diaz, and Tetsuya Sakai. 2013. TREC 2013 Temporal Summarization.. In *TREC*.
[6] Ching-man Au Yeung and Adam Jatowt. 2011. Studying how the past is remembered: towards computational history through large scale text mining. In *CIKM*.
[7] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. *Dbpedia: A nucleus for a web of open data*. Springer.
[8] Steven Bethard. 2013. Cleartk-timeml: A minimalist approach to tempeval 2013. In ** SEM*.
[9] Roi Blanco, Giuseppe Ottaviano, and Edgar Meij. 2015. Fast and Space-Efficient Entity Linking for Queries. In *WSDM*.
[10] Ignacio Cano, Sameer Singh, and Carlos Guestrin. 2014. Distributed non-parametric representations for vital filtering: UW at TREC KBA 2014.
[11] Andrea Ceroni, Ujwal Gadiraju, Jan Matschke, Simon Wingert, and Marco Fisichella. 2016. Where the Event Lies: Predicting Event Occurrence in Textual Documents. In *SIGIR*. ACM.
[12] Jeffrey Dalton, Laura Dietz, and James Allan. 2014. Entity query feature expansion using knowledge base links. In *SIGIR*. ACM.
[13] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2010. A latent variable model for geographic lexical variation. In *EMNLP*.
[14] Paolo Ferragina and Ugo Scaiella. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *CIKM*. ACM.
[15] Goran Glavaš and Jan Šnajder. 2015. Construction and evaluation of event graphs. *Natural Language Engineering* 21, 04 (2015).
[16] Daniel Gomes, João Miranda, and Miguel Costa. 2011. A survey on web archiving initiatives. In *TPDL*. Springer.
[17] David Graus, Maria-Hendrike Peetz, Daan Odijk, Ork de Rooij, and Maarten de Rijke. 2013. yourHistory–Semantic linking for a personalized timeline of historic events. In *Workshop: LinkedUp Challenge at OKCon*.
[18] Dhruv Gupta. 2016. Event Search and Analytics: Detecting Events in Semantically Annotated Corpora for Search and Analytics. In *WSDM*.
[19] Faegheh Hasibi, Krisztian Balog, and Svein Erik Bratsberg. 2016. Exploiting Entity Linking in Queries for Entity Retrieval. In *ICTIR (ICTIR '16)*. ACM, 10.
[20] Helen Hockx-Yu. 2014. Access and scholarly use of web archives. *Alexandria: The Journal of National and International Library and Information Issues* 25, 1-2 (2014).
[21] Susan D Hyde, Nikolay Marinov, and Vera Troeger. 2012. Which elections can be lost? *Political Analysis* (2012), 191–210.
[22] Adam Jatowt and Ching-man Au Yeung. 2011. Extracting collective expectations about the future from large text collections. In *CIKM*.
[23] Chris Kedzie, Kathleen McKeown, and Fernando Diaz. 2014. Summarizing disasters over time. In *Workshop on Social Good (with SIGKDD)*.
[24] Alexander Kotov and ChengXiang Zhai. 2012. Tapping into knowledge base for concept feedback: leveraging conceptnet to improve search results for difficult queries. In *WSDM*.
[25] Erdal Kuzey, Jilles Vreeken, and Gerhard Weikum. 2014. A fresh look on knowledge bases: Distilling named events from news. In *CIKM*.
[26] Jill Lepore. 2015. The Cobweb: Can the Internet be archived? *The New Yorker* (2015).
[27] David Lewis. 1995. The TREC-4 filtering track. In *TREC*.
[28] Hang Li. 2014. Learning to rank for information retrieval and natural language processing. *Synthesis Lectures on Human Language Technologies* 7, 3 (2014).
[29] Xitong Liu and Hui Fang. 2015. Latent entity space: a novel retrieval approach for entity-bearing queries. *Information Retrieval Journal* 18, 6 (2015).
[30] Peter Lyman and Brewster Kahle. 1998. Archiving digital cultural artifacts. *D-Lib* 4, 7 (1998).
[31] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995).
[32] Ian Milligan, Nick Ruest, and Jimmy Lin. 2016. Content selection and curation for web archiving: The gatekeepers vs. the masses. In *JCDL*. IEEE.
[33] Arunav Mishra and Klaus Berberich. 2015. EXPOSE: EXploring Past news fOr Seminal Events. In *WWW*.
[34] Arunav Mishra and Klaus Berberich. 2016. Event digest: A holistic view on past events. In *SIGIR*. ACM.
[35] Federico Nanni, Simone Paolo Ponzetto, and Laura Dietz. 2016. Entity relatedness for retrospective analyses of global events. In *NLP+CSS Workshop at WebSci*.
[36] Alexandros Ntoulas, Junghoo Cho, and Christopher Olston. 2004. What's new on the web?. In *WWW*. ACM.
[37] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation.. In *EMNLP*, Vol. 14.
[38] Jeffrey Pound, Peter Mika, and Hugo Zaragoza. 2010. Ad-hoc object retrieval in the web of data. In *WWW*.
[39] Hadas Raviv, Oren Kurland, and David Carmel. 2016. Document Retrieval Using Entity-Based Language Models. In *SIGIR*.
[40] Petar Ristoski and Heiko Paulheim. 2016. Rdf2vec: Rdf graph embeddings for data mining. In *ISWC*. Springer.
[41] Sylvie Rollason-Cass and Scott Reed. 2015. Living Movements, Living Archives: Selecting and Archiving Web Content During Times of Social Unrest. *New Review of Information Networking* 20, 1-2 (2015).
[42] Michael Schuhmacher, Laura Dietz, and Simone Paolo Ponzetto. 2015. Ranking Entities for Web Queries through Text and Knowledge. In *CIKM*.
[43] Jaspreet Singh, Wolfgang Nejdl, and Avishek Anand. 2016. Expedition: A Time-Aware Exploratory Search System Designed for Scholars. In *SIGIR*. ACM.
[44] Rachele Sprugnoli and Sara Tonelli. 2016. One, no one and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective. *Natural Language Engineering* (2016).
[45] John Tuck. 2008. Web Archiving in the UK: Cooperation, Legislation and Regulation. *Liber Quarterly* 18, 3-4 (2008).
[46] Ian Witten and David Milne. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Workshop on Wikipedia and Artificial Intelligence at AAAI*.
[47] Julian Wolfreys. 2000. *Readings: Acts of close reading in literary theory*. Edinburgh University Press.
[48] Chenyan Xiong and Jamie Callan. 2015. Esdrank: Connecting query and documents through external semi-structured data. In *CIKM*.