Entity Relatedness for Retrospective Analyses of Global Events

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

ABSTRACT

Tracking global events through time would ease many diachronic analyses which are currently carried out manually by social scientists and humanities scholars. While entity linking algorithms can be adapted to identify mentions of an event that goes by a common name, such name is often not established in early stages leading up to the event. This study evaluates the utility of entity relatedness for the task of identifying entities related to the event and textual resources that describe the involvement of the entity in the event. In a small study we find that simple relatedness methods obtain MAP score of 0.74, outperforming many advanced baseline systems such as Stics and Wiki2Vec. A small adaptation of this method provides sufficient explanations of entity involvement on 68% of relevant entities.

1. INTRODUCTION

The World Wide Web contains vast amounts of up-to-date information about nearly every event happened in the world, in the form of blogs, Wikipedia pages, or commentary of experts and eye-witnesses. Moreover, while web archive initiatives are preserving the web for future studies [10], large-scale digitisation projects are constantly expanding this corpus, by making collections of analogue resources (such as newspapers archives [23]) available online. In the last twenty years, the field of Natural Language Processing made progress on extracting information about events (such as conflicts, elections, revolutions, etc.) from text [4], in order to provide a timeline analysis [3], a ranking by "importance" [1] and to automatically enrich knowledge bases [18].

In this work, we focus on "named events", which are events that go by several common names, under which they are identified in text and sometimes even in knowledge bases. In the context of knowledge bases, we refer to named events as "event entities" in the remainder of this paper.

The goal of our work is to support event-centered retrospective analyses, a common research task in social sciences and humanities that is traditionally conducted manually. Consider, for example, the Wall Street Crash of 1929. Historians, sociologists and social economists, interested in understanding how different social groups have experienced it, will start by (1) collecting primary sources that could be generally related to the event (in the form of newspapers articles as well as personal diaries). Then, in a traditional fashion, they will go manually through the documents conducting (2) a meticulous close-reading of each passage and (3) an hermeneutic interpretation of all detected traces, which could be for example mentions of known people and organisation involved in the event, such as William C. Durant or the Rockefeller family. This entire process will help them narrowing down the corpus in a subset of documents closely related to the topic.

Of these steps, especially step (3) requires a vast amount of manual analysis because contextual information surrounding each trace could guide in getting a new perspective on the impact of the overall event. The collection, passage selection, and analysis could theoretically be solved adopting entity linking applied to event entities. However, event entities can only be linked once they go by an established name, such as the Wall Street Crash of 1929 which is also known as the Black Tuesday, the Great Crash, or the Stock Market Crash of 1929. In the crucial early stages leading up to the event, especially for conflicts and revolutions (such as the Gulf War and the Orange Revolution), the event name is not defined or still evolving. As a consequence, an entity linking approach for named events is very likely to suffer from the low recall problem or fail completely when applied to documents from the early stages (as represented in Figure 1). However, such early stages are extremely important when analysing the causes and preconditions that enabled an event.

For these reasons, our research focuses on retrospective analysis of the early stages leading up to the event. As the event name is usually not yet established in that period, our best hopes are to identify other entities (such as people and organisations), which we know are also highly involved in

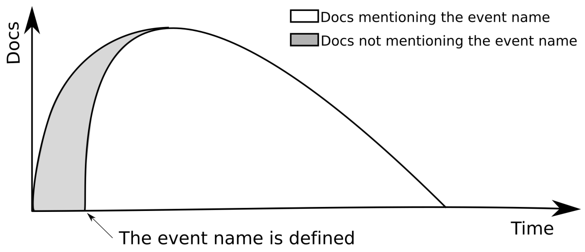

Figure 1: Related documents on an event timeline.

WebSci '16 May 22–25, 2016, Hannover, Germany
© 2016 ACM. ISBN XXXXXXXXXXX…XXXXXXXX
DOI: XXXXXXXXXX

the event and that can help us trace it back through time. Of course, only a small fraction of documents that refer to these central entities, is actually about the event. This is true for all mentions of the entity that precede the beginning of the event or denote other involvements of the entity. Our approach therefore contains two important steps: Extracting event-related entities and detecting text passages that mention the entity in the context of the event. This paper describes ongoing work towards this goal.

Problem Statement. Given a named event in the form of its Wikipedia page title V, predict a ranking of related entities E (Task 1). Furthermore, for each entity E predict a snippet S that explains how the entity is involved in the event (Task 2).

Current entity-based search-tools, such as [15], or Wikipedia's search function already suggest related entities to a specific event. However, to the best of our knowledge, the relative performance of these methods for event-based entity relatedness has not been fully evaluated yet. Additionally, these systems follow the predominant mode of listing related entities as ten hyperlinks in web-search, without presenting additional information on the reason of their relation with the event. As a matter of fact, simply listing a series of related entities makes it extremely hard to gather and synthesise all information required to understand whether a specific entity (e.g. Flaperon) is related to a complex event (e.g., the crash of Malaysia Airlines Flight 370) and does not help a user in expanding her/his understanding. One issue is that the reader may not be aware of the entity, so a short explanation is helpful, the other issue is that entities have many aspects of which only one may be its involvement in the event.

Contributions of this work. In this work we present:

- an evaluation of different methods for identifying related entities, given a relevant event;
- an extensive error analysis that highlights the benefits and weaknesses of each solution;
- an approach for retrieving additional information on the entity's involvement in the event, that could sustain fine-grained event understanding and serve as input for entity-aspect based text classification.

Outline. The paper is organised as follows. First, we present works that are related to our study. Section 3 describes our approach. Section 4 details the dataset, the gold standard, and the results of our experimental evaluation. In Section 5, we highlight benefit and weakness of each evaluated approach before concluding the paper.

2. RELATED WORK

Events and entities. The importance of employing geographical [11] and temporal [17] information in order to gain a better understanding of social phenomena through language is a relevant topic in Natural Language Processing. A large amount of work focuses on detecting stories (such as events) in documents [4], enriching knowledge bases with event-related information [19], associating Wikipedia excerpts describing events to past news articles [21] and combining historical events with information from social media [13]. The task of identifying important events adopting named entities has been recently addressed in [1] and [14]. Entities have also been used to study the general perception of society towards past events [5]. In [18], authors employ named entities to extract yet unknown named events for knowledge base population. Our work moves in the opposite direction, identifying entities related to named events and project them back in time.

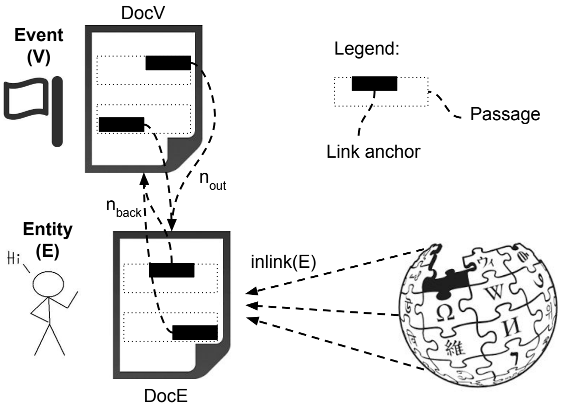

Figure 2: Schema of the adopted method.

Entity Relatedness. The last ten years have witnessed an increased interest in computing relatedness between entities [16]. In [25] it is shown how the hyperlink structure of Wikipedia is an effective low-cost measure of semantic relatedness between Wikipedia articles. Most recent works rely on the the DBpedia knowledge base [6] for acquiring fine-grained information about entities and their semantic relations [24] and for computing document relatedness [22]. In our study we focus on a specific type of entity relatedness, namely the relation between a fixed named event and other entities which are linkable across time.

Passage retrieval. Passage retrieval is often cast as a variation on document retrieval, where the document retrieval model is applied only to a fragment of the text. The applications include search snippet detection, which aims to summarize the query-relevant parts of a document. Scores under the passage model can be combined with those from the containing document to improve performance [8] or to include quality indicators [7]. These approaches have been adapted to retrieve answers for questions [2]. In contrast, this work is set to retrieve passages that explain how an entity is connected to a named event.

3. METHOD

Given an event V, we study two tasks: Identifying related entities E and providing explanatory passages S for each entity.

Preprocessing. For each event, we download the associated Wikipedia article of the event entity V, presented in Figure 2 as DocV. Using the MediaWiki API, we obtain all outlinks for the event page as entity candidates E and download their pages (DocE). In order to obtain positional link information, the text of each article is processed by TagMe! [12]. As no entity linking method is perfect, we discard all entity links produced by TagMe to targets that are not linked according to the MediaWiki API. We obtain statistics of number

Table 1: Wikipedia IDs of the events.

2012-13_Egyptian_protests
2013-14_Thai_political_crisis
2014_Crimean_crisis
2014_Ukrainian_revolution
Charlie_Hebdo_shooting
Ebola_virus_epidemic_in_West_Africa
Global_surveillance_disclosures_(2013-present)
Malaysia_Airlines_Flight_370
Scottish_independence_referendum,_2014
Syrian_Civil_War

inlink(E) pages linking to E provided by DBpedia (Version 04-2015).

Event-Entity Relatedness We combine several link-based entity relatedness methods (following related work on entity relatedness such as [25]) to provide a ranking score for each candidate entity E:

Out Freq Under the assumption that more frequently mentioned entities are more relevant, we include a ranking by the number n_{out} of times the event page $DocV$ contains a link to entity E.

Out TF-IDF In order to bias against generally popular entities, we compute TF-IDF over entities. We do so, by normalizing the outlink frequency by the number of Wikipedia articles that link to the entity E.

Back Freq Suspecting that important entities E will also link back to the event page $DocV$, we include a ranking of E by frequency n_{back} of links back from the DocE to V.

Balance Freq As a measure of symmetry of links, we finally include a ranking by difference of link frequencies: $-|n_{\text{out}} - n_{\text{back}}|$

Each feature provides a ranking over entities, placing entity E at rank r_E. These rankings are aggregated with a simple unsupervised rank-based aggregation as $\sum \frac{1}{r_E}$.

Presenting Explanatory Passages. Given the event V and an entity E, we collect a set of candidate support passages to show which aspects of the entity are relevant, and how this relevance is expressed. As candidate passages we enumerate sentences in V's article that contain a link to E. We combine the first three sentences of this collection and present it to the user.

Given that this method takes inspiration from the system Queripidia [9], we will refer to it in the following sections as "Eventipidia".

4. EXPERIMENTAL EVALUATION

The following experimental evaluation permits us to study the quality of our approaches for event-based entity relatedness and explanatory snippets. As this is our first study in this line of work, we limit it to a small dataset of ten events from the most relevant global events of the recent years. We associate each of these events to its related Wikipedia page. The Wikipedia page ID of each event is presented in Table 1.

Table 2: Results on entity relatedness.

System	MAP@10	Micro-Prec@10
Stics	0.54 ± 0.07	0.59 ± 0.05
Wiki2Vec	0.59 ± 0.11	0.64 ± 0.04
WikipediaRanking	0.66 ± 0.09	0.71 ± 0.05
Eventipedia (our)	**0.74 ± 0.05**	**0.81 ± 0.04**

4.1 Event-based Entity Relatedness

First, we compare our simple approach towards finding event-related entities to three baseline systems using a manually created pool-based gold standard.

Baseline systems. The first method establishes semantic similarities between event and entities adopting word embedding representations [20] of their Wikipedia articles, and ranks entities by similarity. We refer to it as **Wiki2Vec**.[1] The second method adopts the ranking of the search-box provided by Wikipedia to retrieve other entities related to the event-query.[2] We refer to this method as **WikipediaRanking**. As third system we use the entity ranking presented by **Stics** [15] using the event as entity query (not as query terms). It is important to notice that the last system is the only one in our study that does not rely on Wikipedia content information to establish relatedness between events and entities, as Stics collect news from the live web and it links entities to the YAGO knowledge base. In the following section, we will highlight strengths and weaknesses of each system.

Gold standard. For every event V, each system generates a pool of candidate entities E. We asked human annotators to assess these entities for their relevance to the event on a binary scale. In order to support the annotators, we further displayed the explanation snippet (Task 2) or the introduction of the entity's Wikipedia article.

In total, this leads to a set of 629 entity annotations, with 391 entities (62%) being annotated as relevant (two annotators, percentage of agreement 70%). We consider this gold standard dataset as a useful contribution for further research and will make it available with this publication[3].

Results. For each event, the different systems present a list of entities, ranked by their relevance to the event. We evaluate the quality of the ranking using mean average precision at cutoff rank 10 (MAP@10). Additionally we interpret it as a classification task, and report classification performance as micro-averaged precision at 10. The results are presented in Table 2.

In both the interpretation as a ranking task and as a classification task, we observe the same relative performance among the systems, with our Eventipedia system, despite its simplicity, performing best on this event-based entity relatedness task.

4.2 Explanatory Snippets

Finally we conduct a first experimental evaluation of the quality of snippets that explain the involvement of the en-

[1] Wiki2Vec available on https://github.com/idio/wiki2vec

[2] https://en.wikipedia.org/w/index.php?search

[3] Gold standard available at: https://federiconanni.com/2016/04/28/entities-events-relatedness/

Table 3: Results on explanatory passages.

		Eventipedia Snippet			
		Rel.	Non-Rel.	Missing	∑
Wiki-Intro	Relevant	85	10	80	175
	Non-Rel.	180	5	31	216
	∑	265	15	111	

tity in the event. We compare the snippets provided by the Eventipidia system with the first sentences of the entity's Wikipedia article (denoted **Wiki-Intro**).

Gold standard. For all 391 relevant entities among the 10 events, we present both the Eventipedia snippet and the Wiki-Intro to human annotators and ask (separately) whether they explain the involvement of the entity in the event. Assessments are obtained on a binary scale and we ask annotators to read the full Wikipedia article to understand their relation if necessary. In this way we obtain 391 annotations for Wiki-Intro. As the Eventipedia method can only provide snippets when a link from the event to the entity exists, we obtain only 280 annnotations for Eventipedia snippets.

As an explanatory example, we present here the two snippets regarding the relation between the entity Flaperon and the crash of the Malaysia Airlines Flight 370.

> **Wiki-Intro.** A "flaperon" (a portmanteau word) on an aircraft's wing is a type of control surface that combines the functions of both flaps and ailerons. Some smaller kitplanes have flaperons for reasons of simplicity of manufacture, while some large commercial aircraft may have a flaperon between the flaps and aileron.
>
> **Eventipidia.** Nothing was found of the aircraft until 29 July 2015, when a piece of marine debris, later confirmed to be a flaperon from Flight 370, washed ashore on Reunion Island.

Results. In Table 3 we report the results of our analysis as a confusion matrix, providing a separate column for entities where Eventipedia cannot provide a snippet.

We see that in 45% of the cases, the Wiki-Intro was a sufficient explanation. In contrast, the Eventipedia approach provides sufficient explanations in 68% of the cases. The majority of the remaining 32% is attributed to the case where no snippet could be produced. We want to point out that for nearly all cases, where Eventipedia does provide a snippet, this is also relevant. This indicates the our simple method for finding explanations is a reliable source with potential for exploitation of training text-based classifiers for event tracking.

In contrast, the Wiki-Intro only provides a good explanation in 42% of the cases. We suspect that many event-relevant entities are often more popularly known for other accomplishments and therefore the first paragraph is not always a good summary. However, the Wiki-Intro provides a sufficient explanation for most of the missing Eventipedia snippets.

While this study is carried out in a small scale, these results are in line with a previous study on finding entity explanations for web queries [9].

5. ERROR ANALYSIS

In order to have a better understanding of different strengths and weaknesses of the systems, we provide some narrative evaluation. The events on which all systems perform best are the 2014 Ukrainian Revolution and the Global surveillance disclosures. Regarding the first event, all systems are able to identify the importance of entities such as Crimea and Viktor Yanukovych as well as demonstrations such as Euromaidan. Regarding the second event both Edward Snowden and the National Security Agency are detected as extremely related entities, together with other entities such as the surveillance program ECHELON. However, if we consider the most common errors the different systems made, they are substantially different:

Stics relies on a collection of news articles, from which it extracts the most frequent named entities (with a focus on people, locations, and organisations) related to a query. For this reason, we suspect that its knowledge base is narrower than the other solutions, which work directly with Wikipedia. However, the most common mistakes made by Stics depends on the way the collected news have been pre-processed. For example, if we consider the most frequent entities presented by searching the event Syrian Civil War[4] the fourth and sixth entities retrieved are Google and Thomson Reuters. This is probably due to the fact that these entities are not filtered out during the news crawling process. The same issue appears while searching for other events, such as the Charlie Hebdo shooting.

Wiki2Vec is the system that presents the most inconsistent performance. While this system is able to retrieve a series of extremely related entities for the 2012-13 Egyptian protests, such as preceding events (e.g. the Egyptian Revolution of 2011), it also presents a set of completely unrelated entities for the events Scottish independence referendum, 2014 and especially for the Malaysia Airlines Flight 370 disappearance. By looking at the results of this second event, we find a list of other flight disasters or disappearance, such as the Northwest Airlines Flight 255 and the Alaska Airlines Flight 261 where the only relation with the event is the "type" of the event. These examples clearly highlight the potential and the limitations of word-embeddings for the detection of entities related to a specific event. As Wiki2Vec detects semantic similarities between two entities articles, this ranking could be useful for determining sub-events leading up to and following a main event, but it could be misleading when the event is for example a political process that is repeated in time.

WikipediaRanking strongly relies on the previous existence of a specific Wikipedia category for the event. When this category exists, as for example for the Scottish independence referendum, 2014, the retrieved entities are extremely relevant to the event, as they have been initially manually selected by humans. At the same time, when an event category has not yet been created, the search tool computes string matching between the words in the event name and entity articles. As a consequence, for such events like the 2013-14 Thai political crisis, the system retrieves a series of pages that are completely unrelated to the topic, like the Thai general election, 1952.

[4] Experiment conducted on the 23th of March, 2016.

Eventipedia computes event-entity relatedness analysing the hyperlinked structure of Wikipedia. Our study confirms results on standard entity relatedness [25], with MAP@10 performance of more than 0.54 across all events and minimum precision of 0.6. However, as it relies on TagMe! for detecting positional links, recent entities could be missed while processing the text, as TagMe! is based on Wikipedia snapshots of July 2012. Moreover, this approach tends to privilege specific entities over the most commonly mentioned entities. For this reason, while it is able to detect the importance of the entity Flaperon for the event Malaysia Airlines Flight 370, it also presents on top of the ranking a list of French regions such as Île-de-France, where soldiers were deployed for the manhunt following the Charlie Hebdo shooting or the city Perth for the event Malaysia Airlines Flight 370, as the primary search area was identified at about 1,800 kilometres south-west of it.

6. CONCLUSIONS

In this paper we introduce our research, focused on identifying event related entities given a named event as a query. Key to our approach are state-of-the-art methods that measure relatedness between event and entities from the Wikipedia hyperlink structure and a solution for retrieving explanatory passages on the relation from entity links. In a small study we find that simple relatedness methods obtain MAP score of 0.74, outperforming many advanced baseline systems such as Stics and Wiki2Vec. A small adaptation of this method provides sufficient explanations of entity involvement on 68% of relevant entities.

The next step will explore how the information on event-related entities and their explanations can be exploited to track events through time. In particular we hope to use related entities and language of explanations to track the event during early stages where the event does not yet have an established name and therefore is not accessible to an event-based entity linking method.

7. AKNOWLEDGEMENT

This research project is funded through the Eliteprogram for Postdocs by the Baden-Wuerttemberg Stiftung.

8. REFERENCES

[1] A. Abujabal and K. Berberich. Important events in the past, present, and future. In *Proc. of WWW*, 2015.

[2] E. Aktolga, J. Allan, and D. A. Smith. Passage reranking for question answering using syntactic structures and answer types. In *Advances in Information Retrieval*. Springer, 2011.

[3] J. Allan. *Topic detection and tracking: event-based information organization*, volume 12. Springer Science & Business Media, 2012.

[4] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *Proc. of SIGIR*, 1998.

[5] C.-m. Au Yeung and A. Jatowt. Studying how the past is remembered: towards computational history through large scale text mining. In *Proc. of CIKM-11*, 2011.

[6] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. *Dbpedia: A nucleus for a web of open data*. Springer, 2007.

[7] M. Bendersky, W. B. Croft, and Y. Diao. Quality-biased ranking of web documents. In *Proc. of WSDM-11*, 2011.

[8] M. Bendersky and O. Kurland. Utilizing passage-based language models for document retrieval. In *Advances in Information Retrieval*. Springer, 2008.

[9] L. Dietz, M. Schuhmacher, and S. P. Ponzetto. Queripidia: Query-specific wikipedia construction. *Proc. of AKBC-14*, 2014.

[10] M. Dougherty, E. T. Meyer, C. M. Madsen, C. Van den Heuvel, A. Thomas, and S. Wyatt. Researcher engagement with web archives: State of the art. *Joint Information Systems Committee Report*, 2010.

[11] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *Proc. of EMNLP*, 2010.

[12] P. Ferragina and U. Scaiella. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proc. of CIKM-10*, pages 1625–1628. ACM, 2010.

[13] D. Graus, M.-H. Peetz, D. Odijk, O. de Rooij, and M. de Rijke. yourhistory–semantic linking for a personalized timeline of historic events. In *Workshop: LinkedUp Challenge at OKCon*, 2013.

[14] D. Gupta. Event search and analytics: Detecting events in semantically annotated corpora for search and analytics. In *Proc. of WSDM*, 2016.

[15] J. Hoffart, D. Milchevski, and G. Weikum. STICS: Searching with Strings, Things, and Cats. In *Proc. of SIGIR-14*, 2014.

[16] J. Hoffart, S. Seufert, D. B. Nguyen, M. Theobald, and G. Weikum. Kore: keyphrase overlap relatedness for entity disambiguation. In *Proc. of CIKM-12*, 2012.

[17] A. Jatowt and C.-m. Au Yeung. Extracting collective expectations about the future from large text collections. In *Proc. of CIKM-11*, 2011.

[18] E. Kuzey, J. Vreeken, and G. Weikum. A fresh look on knowledge bases: Distilling named events from news. In *Proc. of CIKM-14*, 2014.

[19] E. Kuzey and G. Weikum. Evin: building a knowledge base of events. In *Proc. of WWW*, 2014.

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 2013.

[21] A. Mishra. Linking today's wikipedia and news from the past. In *Proc. of the 7th Workshop for Ph. D Students at CIKM*, 2014.

[22] Y. Ni, Q. K. Xu, F. Cao, Y. Mass, D. Sheinwald, H. J. Zhu, and S. S. Cao. Semantic documents relatedness using concept graph representation. In *Proc. of WSDM*, 2016.

[23] A. Pekárek and M. Willems. The europeana newspapers–a gateway to european newspapers online. In *Progress in Cultural Heritage Preservation*. Springer, 2012.

[24] M. Schuhmacher and S. P. Ponzetto. Knowledge-based graph document modeling. In *Proc. of WSDM*, 2014.

[25] I. Witten and D. Milne. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence*, 2008.