Building Entity-Centric Event Collections For Supporting Research in Political and Social History

Federico Nanni
federico@informatik.uni-mannheim.de
University of Mannheim, Germany

Nikolay Marinov
marinov@sowi.uni-mannheim.de
University of Mannheim, Germany

Simone Paolo Ponzetto
simone@informatik.uni-mannheim.de
University of Mannheim, Germany

Laura Dietz
dietz@cs.unh.edu
University of New Hampshire, United States of America

Introduction

The World Wide Web provides the research community with an unprecedented abundance of primary sources for diachronically tracing, examining and understanding major events and transformations in our society (such as the rise of euroscepticism or the impact of the recent economic crisis). For two decades, public and private institutions have preserved these born-digital materials for future analysis (Gomes and Costa, 2011). However, these collections are now so large that it is infeasible for researchers to study political and social phenomena by examining them in their entirety.

Creating event collections. A common solution that web archives are currently adopting for sustaining the use of the collected sources in humanities research is to offer topic-specific collections. For example, on Archive-it, the Internet Archive presents a few collections on large-scale events such as the Boston Marathon Bombing, Black Lives Matter and the Charlie Hebdo terrorist attack. The collections are curated "by the Archive-It team in conjunction with curators and subject matter experts from institutions around the world" .

Another solution for creating event collections from large datasets is a filtering approach that collects only documents that mention the name of the event; this method has been employed for example in temporal summarization tasks (see Aslam et al., 2013).

Current limitations. The collections created following one of these two approaches share crucial limitations: a) they are small in number; b) the selection process is not always transparent; c) they generally offer only documents that are closely related to an event but lack information on background stories as well as contextual clues. Especially the latter is a crucial issue for historical analyses.

Our vision. We are currently developing a solution for creating event collections that identifies not only the core documents related to the event itself, but most importantly sub-groups of documents which describe related aspects. We do so through an expansion process that is informed by latently relevant concepts and entities from a knowledge base, whose presence in documents is interpreted as one of many indicators of relevance.

Specific contribution. At the DH conference we intend to present the final results of our study, together with its application for supporting research in political and social history.

Method

Let us consider an event, for example the Syrian Civil War, as a node in a knowledge graph (e.g. DBpedia).

As a first step, a domain expert will identify a series of other entities in the knowledge graph that are highly related to the event (in Nanni et al., 2016, we show that this step could be automatized adopting a simple relatedness measure). These could be people, such as Bashar Al-Assad as well as countries (e.g. Turkey, Russia, United States), concepts (e.g. Protests) and other specific events (e.g. The Refugee Crisis). These initial seeds will support us in retrieving other related entities and concepts from the knowledge graph in an automated fashion (we described our solution in Nanni et al. 2016).

While retrieving related entities is important, these are meaningless without human-readable descriptions of the entity's relation to the event. As a matter of fact, the entity United States has many

different aspects, and only few of them are related to the event Syrian Civil War.

In order to retrieve entities in context, we use Wikipedia as an initial corpus. Next, relevant passages from the documents are identified in the collection by information retrieval. Having the entity in context will tell us with which words, concepts and other entities it frequently appears together (a complete overview of the method is presented in Nanni et al., 2017). For example, if a document mentions the United States together with James Foley and ISIS, it is likely to be related to the Syrian Civil War, even without mentioning these words explicitly.

Case studies

We are currently working on two different research tasks:

1. The first study is focused on identifying political speeches on foreign events (such as elections in other countries) in the US Congressional Records (1989-2016), which are available on Congress.gov and through the Internet Archive (Congress.gov provides full-text access to daily congressional record issues dating from 1995, beginning with the 104th Congress. Proceedings for previous years are available on THOMAS). The goal is to measure the amount of attention that US politicians give to international events in correlation with other internal affairs.

2. In the second study, we intend to detect similar patterns during the early rise of anti-establishment protests. Our aim is to uncover small events, which did not turn into large-scale insurrections and therefore are not studied sufficiently. The work is conducted on a large (16 terabyte) web archive of news, blogs, forums and social media, namely the TREC Streaming Corpus (This corpus is a huge web archive collection collected between 2012 and 2014). Finally, the goal of the project is to obtain a better understanding on how and why specific protests succeed while others do not (also in correlation with analyses from the previous study).

Experiments

Identifying related entities. In a previous work (Nanni et al., 2016), we have first established the quality of our entity-relatedness solution (Eventipedia), by comparing it with a series of other baselines commonly used in the field. The results are reported in Figure 1.

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**

Figure 1. Evaluation on entity-event relatedness (from Nanni et al., 2016).

While our entity-relatedness solution outperformed the other tested methods, it also showed a few limitations. In fact, this approach (in its fully automated fashion) tends to privilege specific entities over the most commonly mentioned entities. We address this in developing techniques that are supervised by domain experts; this ensures us to always consider the most relevant related entities.

Collect entities in context. Additionally we studied the identification of human-readable descriptions of the entity's connection to the event. We compared our entity link-based approach with a common information retrieval heuristic which considers the first sentences of the entity's Wikipedia article, as a relevant passage (Wiki-Intro).

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

Figure 2. Evaluation on retrieving entities in context (from Nanni et al., 2016).

The results are presented in Figure 2. In 45% of the cases, the Wiki-Intro was a sufficient explanation. However, our Eventipedia approach provides sufficient explanations in 68% of the cases. We remark that for nearly all cases, where Eventipedia does provide a snippet, this is also relevant. In contrast, the Wiki-Intro only provides a good explanation in 42% of the cases. This is because many event-relevant entities (e.g. the United States) are often more popularly known for other accomplishments and therefore the first paragraph is not a good description of entity involvement in the event.

Retrieve relevant documents. We are currently assessing the quality of our information retrieval solution, which uses entities and contextual passages

to retrieve documents about specific events with an approach based on Dalton et al. (2014). We report here the very first results of our study on retrieving speeches about foreign elections. This work has been conducted both on the US Congressional Records on the New York Times Corpus.

We compared it with two baselines, a) retrieving all documents that mention the name of the country (e.g. "Syria") and b) retrieving all documents that precisely mention the name of the event (e.g. "election in Syria"). It is evident that the first solution is recall-oriented, while the second, already adopted by the TREC temporal summarization task, favors precision.

Given an event, such as the Syrian presidential election, 2007, all three methods produce a ranking of documents. We examine the quality of the ranking considering 15 different elections. For each election, we created a gold standard of 45 documents (relevant and non relevant). Table 1 presents the results in term of mean-average precision (MAP) on the two datasets.

Method	US Congress	NYT Corpus
Place	0.58 ± 0.06	0.48 ± 0.06
EventName	0.32 ± 0.06	0.64 ± 0.06
Eventipedia	0.65 ± 0.06	0.63 ± 0.06

Table 1. First results (MAP) on the collection-building task.

The results of this initial study lead to a few findings. First we see that, especially on the US Congressional Records, using the name of the event permits to collect a fraction of relevant documents, but not all of them. Second, our solution, which uses related entities in context, provides good performance quality on both datasets. Third, our solution is able to identify materials that do not explicitly mention the name of the event.

"Non-relevant" documents. We next analyze what kind of "non-relevant" documents the different systems retrieved among the top elements of the ranking. Non-relevant documents retrieved by the "Event name" solution often discuss different topics and simply mention the event out of context; these documents could be for example general summaries of the previous week.

Instead, non-relevant documents retrieved by Eventipedia are often related to the political activity of a foreign country, but not specifically about the election. For example, they could mention the visit of a candidate to Washington, a few months before the vote. It is evident that choosing one method over the other will shape the event-collection in a different way. Ultimately, it is up to the humanities researcher to decide which documents are most important for the analysis.

Bibliography

Aslam, J. A., et al. (2013) "TREC 2013 Temporal Summarization." *TREC*.

Dalton, J., Dietz, L., and Allan, J. (2014). "Entity query feature expansion using knowledge base links." *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM.

Gomes, D., Miranda, J., and Costa, M. (2011) "A survey on web archiving initiatives." *International Conference on Theory and Practice of Digital Libraries*. Springer Berlin Heidelberg.

Nanni, F., Ponzetto, S. P., and Dietz, L. (2016) "Entity Relatedness for Retrospective Analyses of Global Events", *Proceedings of NLP+CSS: Workshops on Natural Language Processing and Computational Social Science,* at WebSci.

Nanni, F., Ponzetto, S. P., and Dietz, L. (2017, forthcoming) "Building Entity-Centric Event Collections", *Proceedings of the Joint Conference on Digital Libraries,* (forthcoming).