Event Related Document Retrieval with Multilingual Real World Event Representation

Guillaume Bernard[1][0000−0001−5945−4865], Cyrille Suire[1], Cyril Faucher[1], and Antoine Doucet[1][0000−0001−6160−3356]

Université de La Rochelle, Laboratoire L3i, 17000 La Rochelle, France
{guillaume.bernard,cyrille.suire,cyril.faucher,antoine.doucet}@univ-lr.fr
https://l3i.univ-larochelle.fr/

Abstract. This demonstration paper introduces a tool to analyse historical digital libraries with the benefit of publicly available knowledge bases, such as Wikidata and Wikipedia. In this paper, we focus on real-world-events of the past, such as festivals or assassinations. We introduce our method which merges knowledge from Wikidata and Wikipedia article summaries to gather entities involved in events, dates, types and labels. We hereby present the Web tool we designed and the implemented method to characterise events. It integrates easily into any event oriented pipeline: in our demonstration, we use it to find event related documents in the NewsEye corpus [1].

Keywords: Event · Event Representation · Information Retrieval · Linked and Open Data

1 Introduction

We propose a software library [2] able to extract event characteristics from at least two knowledge bases (KB), Wikidata and Wikipedia. Used together, they contain sufficient data to characterise real world historical events. In this paper, we refer to events as *happenings in the real world which have spatio-temporal anchors and additional entities involved in it.* We focus on participating entities which are dates, places and participants [8]. In this paper, we also release a Web demonstrator that is able to build a language-independent, representation of events and to use it in order to query event related documents from a large historical news corpora. For a given language, it provides all the available spellings for the event type and the associated entities. This tool can therefore be useful for improving event-based search engines performance for digital libraries.

[1] The NewsEye corpus is available for anyone at https://platform.newseye.eu

[2] The package is a Python 3 library called `wikivents` on Pypi.org and available on the Software Heritage repository at https://archive.softwareheritage.org/swh:1:dir:ef325a054ba6f7eb1121807da7b1c92b9ecde8f8

With such a tool, we propose to analyse a large corpora of historical documents. The method we propose in this paper will help filtering documents on whether they report a specific historical event. To the best of our knowledge, such a tool does not exist. We propose a method to describe events from knowledge bases, and the interface to use it and find event related documents.

In this paper, we will take the example of the *Assassination of Rasputin* event, a murder that occurred in December 1916 in the Russian Empire.

2 Build a Language-independent Event Representation

Our methodology uses an ontology (*i.e.* Wikidata) as a primary source of information to obtain fundamental event characteristics: type and dates or times. It also processes semi-structured databases (*i.e.* Wikipedia and others) to identify involved entities [8]. We first focused on Wikidata for the links it has with the Wikimedia ecosystem, especially the links between Wikidata entities and Wikipedia articles in all languages.

This software supplies a solution when it is either too long or complicated to aggregate event characteristics manually. It enables to add more data sources (*e.g.* EventKG [3] or DBPedia [2]) to refine the event representation. It can easily integrate a dedicated natural language processing pipeline focused on real-world event analysis. We take the example of researchers wishing to gather all the characteristics about political assassinations provided by a KB. There is a total of 78 entities of this type in Wikidata (this means only 78 Wikidata entries are "instances of" political assassinations), over 952.351 events. Other event types are available, but we focus on what we qualify of indisputable an event that is considered as such for people with various backgrounds (history scholars [7] or NLP reseachers [1, 6], for instance). Political assassinations are an example of an indisputable event type. For each of them, researchers collect the event dates, types and labels from a public API. The added value of this tool is that it is able to gather these information from the ontology and all the entities mentioned in the associated semi-structured databases. EventKG or Wikipedia are not exhaustive databases and some properties, as well as named entities may miss.

2.1 Ontologies: the Extraction of Elementary Event Information

Ontologies such has EventKG [3], YAGO [4] or Wikidata provide different event identifiers which reveal subtleties. In this paper, we conform to the Wikidata event type (WET) definition [5], which includes both Wikidata event types *Q1656682* for breaking events and *Q1190554* for events without premisses.

Our software takes Wikidata entities identifiers as input. *Q2882749* is the input value for our example, the *assassination of Rasputin.* The entity is checked to confirm it is an instance of a WET. If so, the dates (in the Gregorian calendar), the event locations and labels are saved. These generic properties discriminate two similar events: it is unlikely that two distinct events have the same labels, types, and occurred at the same place at the same time. Other properties (the *target* for a *political assassination* for instance) do depend on the WET and are often missing. To overcome this limitation, we propose to analyse Wikipedia lead sections as well, looking for relevant named entities.

In our example, we get the entity labels as strings for every language, the date and the linked entities, identified by their URIs.

2.2 Semi-structured data sources: the Extraction of Entities Involved in the Event

The event representation is supplemented with entities extracted from semi-structured data-sources. In this work, we analyse the lead sections of Wikipedia articles. We focus on those written in five well represented languages that are English, French, German, Spanish and Italian. Languages are chosen arbitrarily from the top-10 list of biggest Wikipedia versions, excluding ones written by bots. For each lead section, involved entities [8] represented as Wikipedia internal links are kept and the corresponding identifiers in the ontology saved.

To show the relevance of entities in relation to the event, the number of occurrences found in lead sections is counted. Entities found in multiple lead sections are important in the event description, synthesise historical knowledge and give an unbiased information about their implication in the event. In our example, there only exists Wikipedia articles written in Spanish and French. From them, we extract, for instance, these triples: *(PER, Q312997 [Felix Yusupov, perpetrator], 3)*, *(PER, Q43989 [Grigori Rasputin, target], 2)*, *(GPE, Q34266 [Russian Empire], 1)*. The weights, respectively 3, 2 and 1 show that knowing who murdered the victim is more pertinent than where it happened. Beside of those entities, the system found 7 people involved, 5 organisations and 4 geo-political entities. There are respectively only 3, 0 and 3 where analysing Wikidata only.

2.3 Event Description in Understandable Languages

The event representation then consists of an association of absolute properties such as dates and links to knowledge bases: it is language independent. In most cases, ontologies provide multiple names in different languages (*i.e.* with different writings) for each entity. In our example, in French the entity *Q312997* is written *Félix Youssoupoff* or *Felix Youssoupov*.

A final step transforms an abstract event representation, based on KB identifiers, into a language-dependent description. It extracts all the alternative spellings for every entity involved in the event. The software makes it possible to get the event description in Italian even if, in this example, only French and Spanish Wikipedias were analysed.

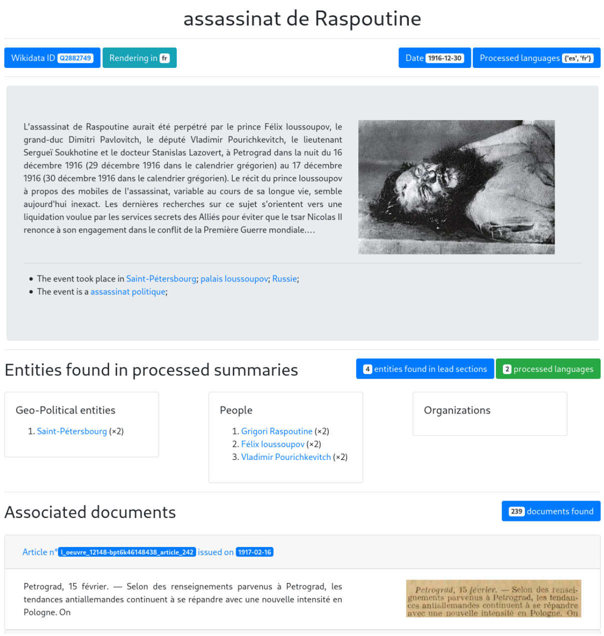
3 Limitations and Opportunities

It may happen that for some entities, there does not exist any language-specific spelling in any ontology (*e.g.* the entity *Q3129997*, Felix Yussopov has no spelling given in some languages). This case is rare but may be encountered when processing events where local entities are involved, less known to speakers of other languages. We only took into account a list of arbitrarily chosen languages, without capitalising on the languages spoken where the event happened, excluding *de facto* Asian and African languages.

The process has drawbacks: SPARQL queries are used to retrieve event data and the procedure can be slow due to the multiple API calls. Our caching implementation accelerates the process by a factor of almost 20 (from one minute to five seconds with the running example). We previously mentioned the possibility to extend the tool capabilities. This way, it is easy to locally process any triple store, such as a Wikidata dump, to overcome this limitation.

Evaluation of such a tool is hard, due to the lack of annotated data. We propose, to evaluate this tool and the efficiency of the event representation to use datasets such as those from the Topic Detection and Tracking (TDT) program [1] or the event-centered dataset from Event Registry [6] which is a collection of news articles, in multiple languages with event annotations.

4 Conclusions and Future Work



assassinat de Raspoutine

Wikidata ID Q2882749 Rendering in fr Date 1916-12-30 Processed languages {'es', 'fr'}

L'assassinat de Raspoutine aurait été perpétré par le prince Félix Ioussoupov, le grand-duc Dimitri Pavlovitch, le député Vladimir Pourichkevitch, le lieutenant Sergueï Soukhotine et le docteur Stanislas Lazovert, à Petrograd dans la nuit du 16 décembre 1916 (29 décembre 1916 dans le calendrier grégorien) au 17 décembre 1916 (30 décembre 1916 dans le calendrier grégorien). Le récit du prince Ioussoupov à propos des mobiles de l'assassinat, variable au cours de sa longue vie, semble aujourd'hui inexact. Les dernières recherches sur ce sujet s'orientent vers une liquidation voulue par les services secrets des Alliés pour éviter que le tsar Nicolas II renonce à son engagement dans le conflit de la Première Guerre mondiale....

- The event took place in Saint-Pétersbourg; palais Ioussoupov; Russie;
- The event is a assassinat politique;

Entities found in processed summaries 4 entities found in lead sections 2 processed languages

Geo-Political entities

1. Saint-Pétersbourg (×2)

People

1. Grigori Raspoutine (×2)
2. Félix Ioussoupov (×2)
3. Vladimir Pourichkevitch (×2)

Organizations

Associated documents 239 documents found

Article n° l_oeuvre_12148-bpt6k46148438_article_242 issued on 1917-02-16

Petrograd, 15 février. — Selon des renseignements parvenus à Petrograd, les tendances antiallemandes continuent à se répandre avec une nouvelle intensité en Pologne. On

Petrograd, 15 février. — Selon des renseignements parvenus à Petrograd, les tendances antiallemandes continuent à se répandre avec une nouvelle intensité en Pologne. On

Fig. 1. Web Event exporter, displaying the Assassination of Rasputine in French

In this paper, we release a set of tools, including a Web front-end to the library, able to gather into one specific language, all the event related knowledge that exists. We provide an easy way to analyse events described on the Internet. The library can be easily extended and integrating more data sources (*i.e.* ontologies) is made easy. Extensive technical information may be found on the Pypi project page, as well as real world examples and outputs.

It is already in use as the entry point of a pipeline in a multilingual historical event based search engine which indexes millions of historical documents. For this demonstration, we used a private access to the NewsEye corpus to build queries with information collected from Wikidata and Wikipedia. Similarly to a search engine query, we weighted the entities according to their importance related to the event. The query is a conjunction of all weighted terms found by the `wikivents` library.

5 Acknowledgments

This work has been supported by the European Union's Horizon 2020 research and innovation program under grants 770299 (NewsEye) and 825153 (Embeddia).

References

1. Allan, J.: Introduction to Topic Detection and Tracking. In: Topic Detection And Tracking: Event-Based Information Organization, pp. 1–16 (2002)
2. Färber, M., Bartscherer, F., Menne, C., Rettinger, A.: Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. Semantic Web **9**(1), 77–129 (Nov 2017). https://doi.org/10.3233/SW-170275
3. Gottschalk, S., Demidova, E.: EventKG: A Multilingual Event-Centric Temporal Knowledge Graph. Lecture Notes in Computer Science **10843** (Jun 2018). https://doi.org/10.1007/978-3-319-93417-$4_1$8
4. Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artificial Intelligence **194**, 28–61 (Jan 2013). https://doi.org/10.1016/j.artint.2012.06.001
5. Rudnik, C., Ehrhart, T., Ferret, O., Teyssou, D., Troncy, R., Tannier, X.: Searching News Articles Using an Event Knowledge Graph Leveraged by Wikidata. In: Companion Proceedings of The 2019 World Wide Web Conference on - WWW '19 (2019). https://doi.org/10.1145/3308560.3316761
6. Rupnik, J., Muhic, A., Leban, G., Skraba, P., Fortuna, B., Grobelnik, M.: News Across Languages - Cross-Lingual Document Similarity and Event Tracking. Journal of Artificial Intelligence Research **55**, 283–316 (Jan 2016). https://doi.org/10.1613/jair.4780
7. Shaw, R.: A Semantic Tool for Historical Events. In: Proceedings of the The 1st Workshop on EVENTS: Definition, Detection, Coreference, and Representation. pp. 38–46. Atlanta, Georgia, USA (Jun 2013)
8. Yadav, V., Bethard, S.: A survey on recent advances in named entity recognition from deep learning models. In: Proceedings of the 27th International Conference on Computational Linguistics (Aug 2018), `https://www.aclweb.org/anthology/C18-1182`