Searching News Articles Using an Event Knowledge Graph Leveraged by Wikidata

Charlotte Rudnik, Thibault Ehrhart, Olivier Ferret, Denis Teyssou, Raphaël Troncy, Xavier Tannier

Searching News Articles Using an Event Knowledge Graph Leveraged by Wikidata

Charlotte Rudnik
LIMSI, CNRS
Orsay, France
charlotte.rudnik@limsi.fr

Thibault Ehrhart
EURECOM
Sophia Antipolis, France
thibault.ehrhart@eurecom.fr

Olivier Ferret
CEA, LIST,
Gif-sur-Yvette, F-91191, France
olivier.ferret@cea.fr

Denis Teyssou
Agence France-Presse (AFP)
Paris, France
denis.teyssou@afp.com

Raphaël Troncy
EURECOM
Sophia Antipolis, France
raphael.troncy@eurecom.fr

Xavier Tannier
Sorbonne Université, Inserm, LIMICS
Paris, France
xavier.tannier@sorbonne-universite.fr

ABSTRACT

News agencies produce thousands of multimedia stories describing events happening in the world that are either scheduled such as sports competitions, political summits and elections, or breaking events such as military conflicts, terrorist attacks, natural disasters, etc. When writing up those stories, journalists refer to contextual background and to compare with past similar events. However, searching for precise facts described in stories is hard. In this paper, we propose a general method that leverages the Wikidata knowledge base to produce semantic annotations of news articles. Next, we describe a semantic search engine that supports both keyword based search in news articles and structured data search providing filters for properties belonging to specific event schemas that are automatically inferred.

ACM Reference Format:
Charlotte Rudnik, Thibault Ehrhart, Olivier Ferret, Denis Teyssou, Raphaël Troncy, and Xavier Tannier. 2019. Searching News Articles Using an Event Knowledge Graph Leveraged by Wikidata. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316761

1 INTRODUCTION

Information and communication technologies have provided tools and methods to make the production of information more democratic. In this context, journalists and technologists have developed the notion of "data journalism", which takes advantage of structured and numerical data in the production and distribution of news. It also takes advantage of the growing popularity of Linked and Open Data and the development of structured knowledge bases such as DBpedia [5], YAGO [3] or Wikidata [17] to facilitate information analysis and to access a variety of points of view. However, knowledge is still far from being entirely represented in structured databases, and the most prominent way to convey information to the end user is still the free text, complemented by multimedia content.

Intertwining structured and unstructured data in information systems is still an open research problem. In this paper, we present a system for aggregating unstructured news articles and structured data describing events leveraging on the Wikidata knowledge base. This approach makes use of several Information Retrieval and Information Extraction tasks. In the context of Information Extraction, the knowledge associated with news articles can typically be used for training event extractors in a distant supervision mode [12]. From the Information Retrieval perspective, the approach makes it possible to retrieve news articles describing events using either keyword-based queries or filters that typically make use of properties available in knowledge bases. It also allows to query Wikidata and then to read an entire annotated article describing the corresponding event. We implemented a system which is available at http://asrael.eurecom.fr/ and covers the last two tasks.

Figure 1 illustrates the architecture of our system. Events described in news articles are mapped to events from Wikidata (Section 4.1), and attributes from the Wikidata instances are used to annotate the news articles when possible (Section 4.3). Wikidata events belong to specific classes, but these classes are too fine-grained for being used in a search engine. Furthermore, many event classes actually share a similar structure (*i.e.* sets of attributes). For example, a general *election* schema is relevant for describing any type of elections regardless of the more specific Wikidata types such as "*Bundestag election*" (`Q1007356`), "*direct election*" (`Q1196727`), "*Elections in Saudi Arabia*" (`Q4119635`)... For these reasons, we add a hierarchical clustering step (Section 4.2) to automatically create coarser grained schemas. Finally, we implemented an event-oriented knowledge graph and a search engine able to query and navigate through both the knowledge base and the news articles (Section 4.4).

2 RELATED WORK

Contrarily to WikiNews[1], Wikipedia does not aim to be a news service. However, Wikipedia's Current Events portal (WCEP[2]) provides a set of pages where primarily events but also trends and developments are listed on a daily basis with links to reference

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316761

[1] https://www.wikinews.org/
[2] https://en.wikipedia.org/wiki/Portal:Current_events

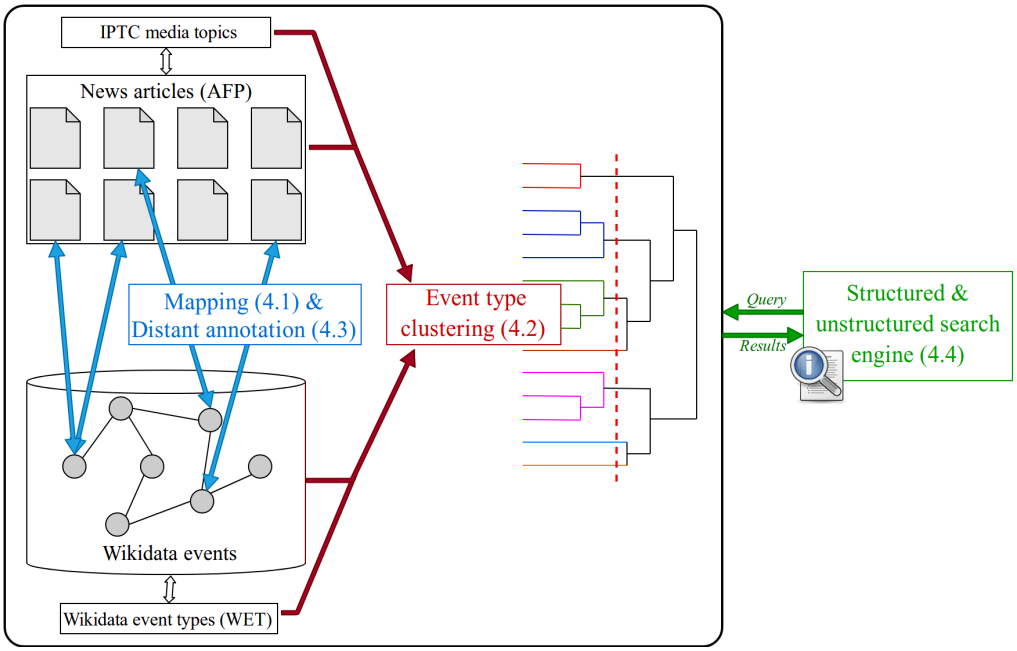

Figure 1: System overview for annotating news articles and enabling structured search.

articles. The WikiTimes project[3] is the first attempt to build a structured and rich knowledge base of news events by harvesting the efforts of the Wikipedia crowd in maintaining WCEP [14]. The WikiTimes knowledge base is represented in RDF and it contains very short descriptions of events that can be filtered by entities, location and time.

Based on this experience, Gottschalk and Demidova have developed the EventsKG knowledge graph[4], a multilingual resource incorporating event-centric information extracted from Wikidata, DBpedia and YAGO, as well as less structured sources such as the Wikipedia Current Events Portal and Wikipedia event lists in five languages [2]. EventsKG re-uses the Simple Event Model (SEM) ontology [16] to describe nearly 700,000 events. However, temporal information is available for 76% of those events and location information for only 12% of them. While many entities are mentioned, extracted and disambiguated in the short descriptions of those events, this is far from being complete. Events are generally weakly categorized and both categorical and numerical data representing the events attributes are rarely extracted. Annotating semantically newsfeeds at scale is being continuously proposed in [13] which maintains the Newsfeed[5] service. Annotations are, however, restricted to named entities that can be extracted by the Enrycher tool[6].

For easing the exchange of news, the International Press Telecommunication Council (IPTC) has developed the NewsML Architecture (NAR), specialized into a number of languages such as NewsML G2 and EventsML G2. As part of this architecture, specific controlled vocabularies, such as the IPTC Media Topics or News Codes, are used to categorize news items together with other industry-standard thesauri. In previous work, we designed an OWL ontology for the IPTC News Architecture and we converted the IPTC NewsCodes into a SKOS thesaurus [15]. IPTC is now publishing itself the IPTC Media Topics in SKOS and has further developed the rNews vocabulary, largely based on Schema.org, for describing news articles. In this work, we re-use the rNews vocabulary to describe the original metadata attached to news articles. Furthermore, we annotate the news articles using properties and entities from Wikidata once events reported in the news have been mapped to existing Wikidata events.

While mapping text to knowledge has been the subject of a large body of work, represented in the recent ages with work such as [7] or all the work about entity linking [1], mapping text to event representations and more particularly news articles to event representations has not been the focus of lots of studies. One exception is [9], followed by [10], which tackles this kind of mapping according to an Information Retrieval perspective through two tasks based on the notion of *Wiki-excerpt*. A *Wiki-excerpt* corresponds to a description of an event built from Wikipedia and contains both a textual description and factual information about the event such as temporal expressions, geolocation and named entities. The first task, *Wiki2News*, starts from a *Wiki-excerpt* and aims at retrieving a set of past news articles about the considered event while *News2Wiki* is the reverse task, consisting in retrieving *Wiki-excerpts* from a set of news articles. The work focuses more specifically on the *Wiki2News* task by designing time-aware language models for supporting the retrieval of past news articles. More recently, the *Wiki2News* task of [9] has been considered under the perspective of the enrichment

[3]http://wikitimes.l3s.de/
[4]http://eventkg.l3s.uni-hannover.de/
[5]http://newsfeed.ijs.si/
[6]http://enrycher.ijs.si/

of Wikipedia from a stream of news by [6]. This work first builds a temporal event chain of the news articles related to the target event and then selects a subset of them according to various representativeness criteria exploited in a learning-to-rank framework. While all the work we have mentioned was based on Wikipedia, our work tackles the *Wiki2News* task by relying on Wikidata as a knowledge base, with a much simpler, still effective, approach.

3 DATA

3.1 AFP News Articles

In this work, we make use of a very large corpus of text newswire written in English provided by the French news agency AFP. More precisely, we use over 2 million articles covering the period 2004-2017. The topics are worldwide news ranging from politics, diplomacy, sports to natural disasters or economy and business. Each document is an XML file compliant with the NewsML standard, containing a title, a document creation time (DCT), a dateline where the article was written, one or several IPTC Media Topics and a set of keywords (slugs), as well as a textual content split into paragraphs.

The main topic of an article is generally a specific event, and sometimes, other older events are referred to in order to look at the current one from a wider perspective. This is why we consider that it is possible to associate an AFP article with one single event. Furthermore, we assume that the title and the first paragraph (lead) describe the event associated with the document. This is a realistic hypothesis since the basic rules of journalism impose that the first sentence should summarize the event by informing on the "5 Ws" (*What*, *Who*, *When*, *Where*, *Why*).

3.2 Wikidata Occurrences

Wikidata is a free and open knowledge base that can be read and edited by both humans and machines [17]. It acts as a central storage for the structured data of its Wikimedia sister projects including Wikipedia, Wikivoyage, WikiNews, and others. In our work, we focus on newsworthy events, *i.e.* something that happens within a locality and a temporality, and that could be described in one or more news articles. In regard to this definition, there are many general and specific classes in Wikidata that are related to events, but all these classes have in common the same parent class named "Occurrence" (`Q1190554`). For example, the event "Cargolux Flight 7933" (`Q3107014`) is an *instance_of* (`P31`) an "aviation accident" (`Q744913`) which is a *subclass_of* (`P279`) an "aviation occurrence" (`Q15733640`) which is a *subclass_of* (`P279`) "occurrence" (`Q1190554`).

In the remainder of this paper, we call Wikidata Event Type (or WET) the value of the property *instance_of* (`P31`) of a Wikidata event. In the previous example, "aviation accident" (`Q744913`) is a WET. An instance can have several Wikidata Event Types.

We lined up on the temporal coverage of the AFP articles corpus and we considered all Wikidata event instances during the period from 2004 to 2017. The event date can be represented by three properties in Wikidata:

- `P585`:*point_in_time* if the event is a one-off event;
- `P580`:*start_time* and `P582`:*end_time* if the event has a duration.

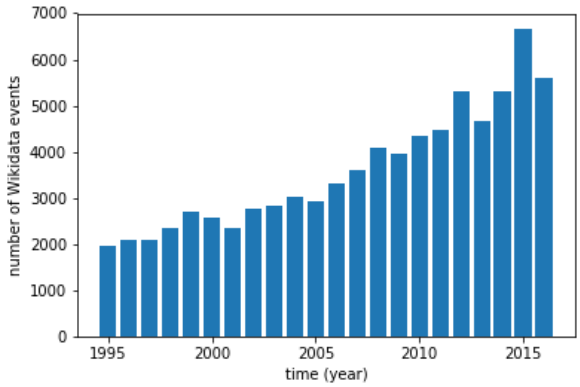

Figure 2: Evolution of the number of Wikidata events over time.

In the second case, we consider the events that have started and ended during the period 2004-2017.

The selection of Wikidata events consists of 60k Wikidata instances. As shown in Figure 2, the distribution over the years is not uniform. The increase in the last few years is explained by a better quality of the data and particularly by the presence of a date in the description of the events, but should not be interpreted as an increase of the number of events happening in the world.

3.3 IPTC Media Topics

The International Press Telecommunications Council (IPTC) maintains a taxonomy of Media Topics[7], which can be seen as a controlled and hierarchical set of indexing keywords. Each article written by AFP is associated with at least one IPTC code by its author. IPTC Media Topics (later called IMTs) give information about the topic of the article and are often linked with a type of event (earthquake, election, crash…), but not always (politics, theatre…). Figure 3 illustrates a part of the IPTC Media Topic hierarchy.

4 APPROACH

Following the description of the data in Section 3, we define that "AFP article" stands for the main event associated with the AFP article and described in the lead, while "Wikidata event" stands for the structured events described in our selection of Wikidata instances. The term 'article' will refer to the news article whereas the term instance will refer to a Wikidata instance.

4.1 Mapping AFP with Wikidata

In order to map Wikidata events to AFP articles, we consider that two mentions of the same event share the following characteristics: same time, same place and same type or category (election, natural disaster, etc.), which we defined as a content similarity.

4.1.1 Scoring function. For mapping a Wikidata instance to an AFP article, we define the following criteria:

[7] http://cv.iptc.org/newscodes/mediatopic, http://show.newscodes.org/index.html?newscodes=medtop

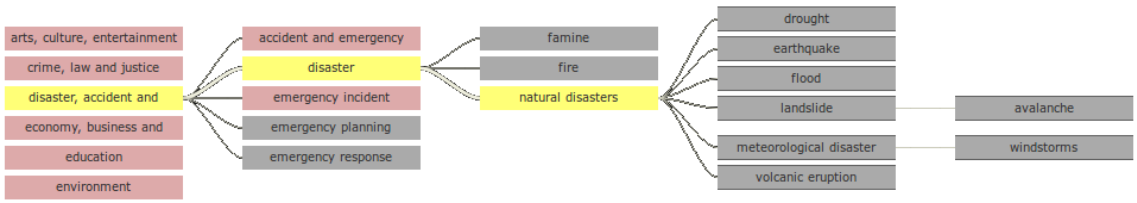

Figure 3: Excerpt of the IPTC Media Topics taxonomy.

Date. The article has to be written at most the day following the *point_in_time*, or between *start_time* and *end_time* if the event has a duration.

Location. The location of a Wikidata instance is defined by the properties *country* (P17) and *location* (P276). One of these values must have be mentioned in the AFP article.

Subject. The Wikidata Event Type (WET) and the title of the instance define a list of keywords relevant to the subject of the article. The similarity score is then the sum of a IMT-based *tf.idf* of the keywords occurring also in the AFP article:

$$score(article) = \sum_{t_i \in article} tf.idf_{IMT}(t_i) \times \mathbf{1}_{keywords}(t_i) \quad (1)$$

where

$$\mathbf{1}_{keywords}(t_i) = \begin{cases} 1 & \text{if } t_i \in WET \\ 0 & \text{otherwise} \end{cases}$$

and

$$tf.idf_{IMT}(t_i) = tf_{IMT}(t_i) \times log\frac{N}{df_{IMT}(t_i)} \quad (2)$$

with t_i the i^{th} token in the article, $df_{IMT}(t_i)$ the number of IMTs associated with documents containing t_i, $tf_{IMT}(t_i)$ the number of occurrences of t_i in the articles sharing the same IMT, and N the number of IMTs. If the AFP article has more than one IMT, then the highest tf_{IMT} for each word is considered.

The weights are based on IPTC Media Topics. Hence, they increase the relevance of a token in a particular context. For instance, "Police" will be more relevant in an event regarding a crime compared to an event regarding an earthquake, even if the token is present in both articles.

An event is considered to be about the same subject as a Wikidata instance if this score is higher than a threshold. Empirically, this threshold is set to 0.04 and roughly corresponds to two tokens in common between the article and the list of keywords.

4.1.2 Evaluation. After this processing, 97,160 articles have been mapped to 8,350 Wikidata instances from 933 Wikidata event types (out of 42k Wikidata events in the considered temporal range). Note that it is neither necessary nor desirable that all articles be mapped to Wikidata instances. Most AFP articles do not relate to events that are meant to end up in Wikidata (*i.e.*, a political speech, a reaction to an event or a trivial news event) and a lot of Wikidata events are not described by any press agency (book publication, local festival or cultural event, TED talks…).

# sentences	Precision	Recall	F1-score
3	1.00	0.67	0.80
5	0.99	0.71	0.83
all	0.96	0.75	0.84

Table 1: Evaluation of the mapping between AFP articles and Wikidata events.

In order to evaluate the quality of this mapping, we manually built a set of 406 pairs *(article, Wikidata instance)* based on 88 instances explicitly linked to a WikiNews page[8] in Wikidata. Only a very small part of Wikidata events are linked to a WikiNews page[9] but we used them to ensure that only "media-worthy" Wikidata events are considered in our manual evaluation. We share this gold standard for further comparison[10].

We report in Table 1 the precision (correct mapping rate), recall (1−missed mapping rate) and F1-measure, when considering the first 3, the first 5 or all the sentences in the AFP articles. These scores show the high quality of the mapping, and consequently, only a very tiny fraction of incorrect information will be shown to the user. Note that missed mappings do not prevent the event to be queried and visualized by our search engine. We only miss the link between the structured and the unstructured data.

4.2 Schema Clustering

As depicted in Table 2, Wikidata event types are often fine-grained and the subclass hierarchy can vary in terms of quality and depth. We seek more coarse-grained event categories for the relevance and robustness of our news classification and potential filters for our search engine. Indeed, from a human perspective, several WETs (e.g. NATO summits and G20 summits) share the same or a very similar structure and clustering them together will make the classification process easier, as well as simplify the interaction with the user within the search engine interface.

To do so, we adopted a hierarchical clustering method based on 3 similarity measures. Each similarity is based on a different representation of the Wikidata Event Type.

- **Label representation**: Even if composed of only a few words (see Table 2), the labels of the Wikidata Event Type (WET) can be a good clue for deciding whether two clusters are similar or not. For instance, the labels "*Election in UK*",

[8] https://en.wikinews.org
[9] Note that our automatic mapping could be used to automate this Wikidata/WikiNews mapping.
[10] https://github.com/crudnik/asrael

Target schema	Related Wikidata event types
Election	Bundestag election, direct election, Elections in Saudi Arabia, ...
Plane crash	Aviation accident, plane crash, mid-air collision, ...
Summit	NATO summit, G20 summit, ...

Table 2: Examples of Wikidata Event Types for three target schemas.

NER label	New mentions
GPE	geopolitical entity
ORG	organization
PERSON	person
NORP	nationality
DATE	date

Table 3: Generic tokens.

"*Election in France*" and "*Election*" should be clustered together. However, this approach needs to exclude mentions of organizations or locations, as in "*earthquake in New Zealand*" compared to "*New Zealand general election*", which are arguably not similar in terms of event type. Therefore, we replace mentions of named entities with generic tokens using the spaCy[11] named entity recognition system following the Table 3 (e.g. "*France*" is replaced by "*GEOPOLITICAL ENTITY*", "*jan-7*" by "*DATE*").

The representation of the labels $R_l(T)$ of a WET T is the mean of the word2vec [8] vectors of their words.

$$R_l(T) = \underset{w \in label(T)}{mean}(w2v(w)) \tag{3}$$

where $label(T)$ is the set of words in T's label and $w2v(w)$ is the word2vec representation of the word w.

- **Content representation**: this representation is based on the content of the articles. A *WET document* is built by concatenating all the articles mapped to this WET at the previous step (Section 4.1). The content representation $R_c(T)$ is then a vector of all words t_i weighted by their $tf.idf_{WET}$, computed as follows:

$$tf.idf_{WET}(t_i) = tf_{WET}(t_i) \times log\frac{M}{df_{WET}(t_i)} \tag{4}$$

where $tf_{WET}(t_i)$ is the number of occurrences of the term in the WET document, $df_{WET}(t_i)$ is the number of WET documents containing the term, and M is the total number of $WETs$ in our dataset.

- **IPTC Media Topic representation**: in order to improve and to facilitate the stopping decision of the clustering, we add a feature based on the IMTs (see Section 3.3).

  As each AFP article is associated with one or several IMTs, the mapping described in Section 4.1 provides also a mapping between a WET and a list of IMTs. We interpret these codes as a new vocabulary describing the WET and we use again a *tf.idf* representation of this new vocabulary. The representation $R_{imt}(T)$ is a sparse vector of size equal to the total number of IMTs present in the corpus, where, for each IMT imt:

$$tf.idf_T(imt) = tf_T(imt) \times log\frac{M}{df_T(imt)} \tag{5}$$

  where $tf_T(imt)$ is the number of articles labeled by the IMT imt that have been mapped to a Wikidata event of type T, $df_T(imt)$ is the number of WETs mapped with at least one article with label imt, M is the total number of $WETs$ in our dataset.

We use these three representations to compute the following similarity between two WETs T_i and T_j:

$$sim(T_i, T_j) = \alpha \times \cos(R_l(T_i), R_l(T_j)) \tag{6}$$
$$+ \beta \times \cos(R_c(T_i), R_c(T_j)) \tag{7}$$
$$+ \gamma \times \cos(R_{imt}(T_i), R_{imt}(T_j)) \tag{8}$$

where cos is the cosine similarity measure and the weights α, β and γ are empirically set as 0.38, 0.57, 0.05 respectively.

Note that each of the "label", "content" and "IMT" representations has a different role. Increasing α gives a higher weight to very short texts, which are generally difficult to compare [4]. This would, for example, make closer labels such as 'strike', 'general strike' and 'military strike', which would decrease the quality of the clustering. Content representation is based on longer pieces of text, but increasing β would give a higher weight to potential errors in the mapping. Finally, the IMT similarity is quite categorical compared to the other ones. We want to use it only as an help for choosing when to stop the clustering, which explains the low value of γ in the global similarity score. The importance of this balance between all sources of information is at the same time a strength and a limitation of the method.

Our agglomerative hierarchical clustering procedure is based on the Ward's method. To cut the resulting dendrogram we used a threshold defined by the elbow method, aiming at finding at which number of clusters the marginal gain of variance will start dropping. According to the graph of Figure 4, the best threshold is 0.23. This leads from 933 initial Wikidata event types to 119 clusters in total. Some extracts of the dendrogram are available in Figure 7.

We observe that this clustering step enables to group together natural disasters, or summits (NATO, G8, ...) into coherent clusters. We also obtained several Election clusters, with the three main ones which seem related to legislative elections, parliamentary elections and general elections.

We empirically evaluated the quality of the clusters for choosing the parameters of our model. Building a protocol for a formal evaluation of this step is a future work. As for most clustering tasks, there is no unique good solution and an automatic, reproducible evaluation seems difficult to set up.

4.3 Automatic Semantic Annotation of News Articles

Our objective is to annotate semantically AFP news articles leveraging Wikidata structured data describing events being told in those

[11] http://spacy.io/

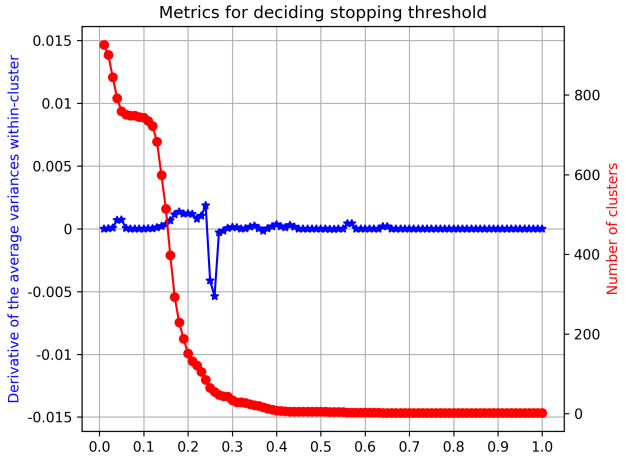

Figure 4: Number of clusters and derivative of the distance within clusters as a function of the similarity threshold to cut the dendrogram.

articles. We distinguish Wikidata properties that have textual values from the ones having numerical values.

4.3.1 Entity Annotation. When Wikidata properties have a textual value, our goal is to search whether this value is present or not in the news article. We also use DBpedia redirections to collect different variations of the mention to search in the text.

4.3.2 Quantity Annotation. When Wikidata properties have a numerical quantity, the problem of finding this information in the news article is much harder. The news articles come as a continuous stream and some reported information generally evolves over time, such as the death toll after a deadly accident. Consequently, a proper semantic annotation of a news article should not always consider the exact quantity value indicated in Wikidata. Therefore, for quantitative values, we introduce some flexibility and a quantity in the text (float or integer) is added to a candidate list of annotations if it is in the +/- 10% range of a property and if it is in the first five sentences of the article (supposedly, the earlier in the text, the most relevant to the article main subject). These candidates are then ranked according to the relevance of the semantic context of the numerical value in the text (considering that the context expresses the type of the attribute). Consequently, given two quantities, the more the context (e.g. "9 were killed on Saturday") is similar with the property type (e.g. *number_of_killed* rather than *magnitude_on_Richter_scale*), the more probable the quantity is to be linked with this property. The news article is then annotated with the wikidata property and the most relevant quantity.

4.3.3 Serializing the Annotations. We represent both the news articles metadata and the semantic annotations in RDF. We first convert the AFP news article metadata encoded in NewsML in RDF using the rNews vocabulary. For example, the metadata associated with the news article described by Listing 1 indicates that this story was created on 24/03/2015 with the English headline *'No survivors' in Germanwings crash: transport minister.*

```
<http://asrael.eurecom.fr/news/71e6c1b5-cbfa-3f85-8510-e200652f6735>
  a  rnews:Article ;
  rnews:dateCreated "2015-03-24T12:41:21Z"^^xsd:dateTime;
  rnews:headline  "'No survivors' in Germanwings crash: transport minister"@en ;
  dc:subject  <http://cv.iptc.org/newscodes/subjectcode/03013000>,
    <http://cv.iptc.org/newscodes/subjectcode/04015000>,
    <http://cv.iptc.org/newscodes/subjectcode/03010000>,
    <http://cv.iptc.org/newscodes/subjectcode/04000000>,
    <http://cv.iptc.org/newscodes/subjectcode/03010003>,
    <http://cv.iptc.org/newscodes/subjectcode/04015001>,
    <http://cv.iptc.org/newscodes/subjectcode/03000000>;
  schema:keywords   "minister", "aviation", "accident",
     "Germany", "Spain", "survivors", "France" .
```

Listing 1: Semantic annotation of a news article

The event being described in this news article exists in Wikidata as `Q19671417`. We create an instance of the `schema:Event`[12] which is about this news article. In this article, the number of dead people (150) is correctly found (Listing 2). The schema *S34* is one of the schema output of the clustering phase described in the section 4.2.

```
<http://asrael.eurecom.fr/news/71e6c1b5-cbfa-3f85-8510-e200652f6735>
   rnews:about  <http://asrael.eurecom.fr/event/71e6c1b5-cbfa-3f85-8510-
         e200652f6735> .

<http://asrael.eurecom.fr/event/71e6c1b5-cbfa-3f85-8510-e200652f6735>
   a    schema:Event , wd:Q750215 , rnews:Concept ;
   rdfs:label  "'No survivors' in Germanwings crash: transport minister" ;
   dc:identifier "urn:newsml:afp.com:20150324T124135Z:TX-PAR-ENS90:5" ;
   owl:sameAs wd:Q19671417 ;
   wdt:P1120  "150" ;
   wdt:schema "S34" .
```

Listing 2: Semantic annotation of a news article

4.3.4 Annotation Dataset. As a result, we created annotations associated with 370 properties extracted from Wikidata. This dataset can easily be used for a relation extraction task, with a distant supervision system. The evaluation of this automatic annotation is part of our future work. Note that this step is not necessary to build the search engine described hereafter.

4.4 Search Engine

We load all RDF annotations in a triple store using the Openlink Virtuoso software. The full text of the news article is also indexed in the triple store. We then developed a user interface that performs SPARQL queries to provide views on the data. The Figure 5 depicts the view of the news article[13]. On the top right, we show an infobox composed of the main named entities extracted in the article using the ADEL system [11]. The Figure 6 depicts the view of the news article[14]. On the left panel, the user has selected the schema *S34* corresponding to crash accident. Therefore, a set of additional properties are automatically suggested as new filters, such as the number of victims. The user has entered the value 50 and the search engines retrieves the news articles that describe crash accidents that have yielded at least this number of victims.

5 CONCLUSION AND FUTURE WORK

In this paper, we develop an event-based search engine capable to query both the structure data of knowledge bases and the unstructured textual content of news articles. This facilitates the navigation through events of the same type and aggregate complementary information about the same event. Furthermore, we produced a semi-automatically annotated text dataset. This dataset could be

[12] The `schema` prefix refers to the Schema.org vocabulary.
[13] http://asrael.eurecom.fr/home/details/7733fcef-feef-3f61-b6af-867298d127fc
[14] http://asrael.eurecom.fr/home/details/71e6c1b5-cbfa-3f85-8510-e200652f6735

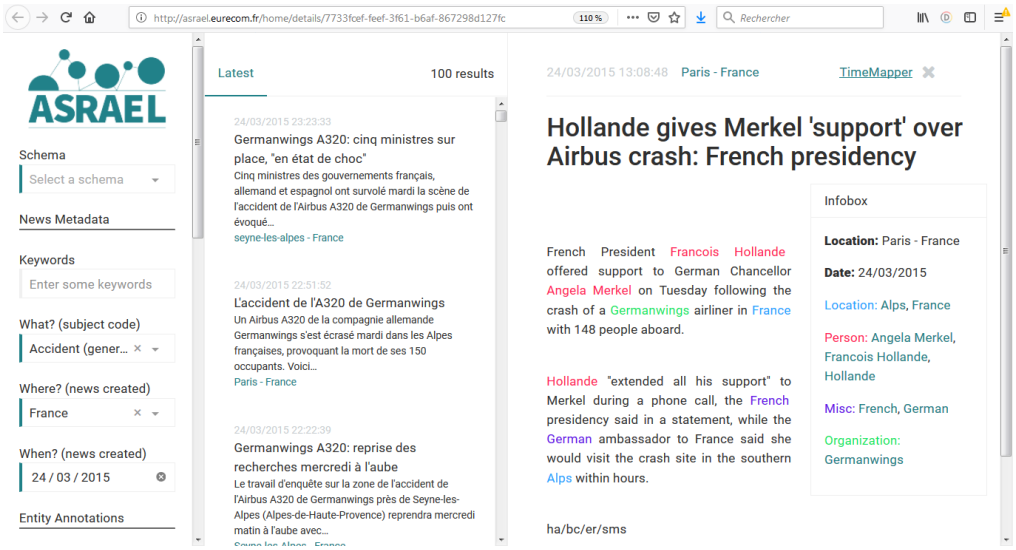

Figure 5: Search engine filtering articles describing plane crash events occurring in France on 24 March 2015.

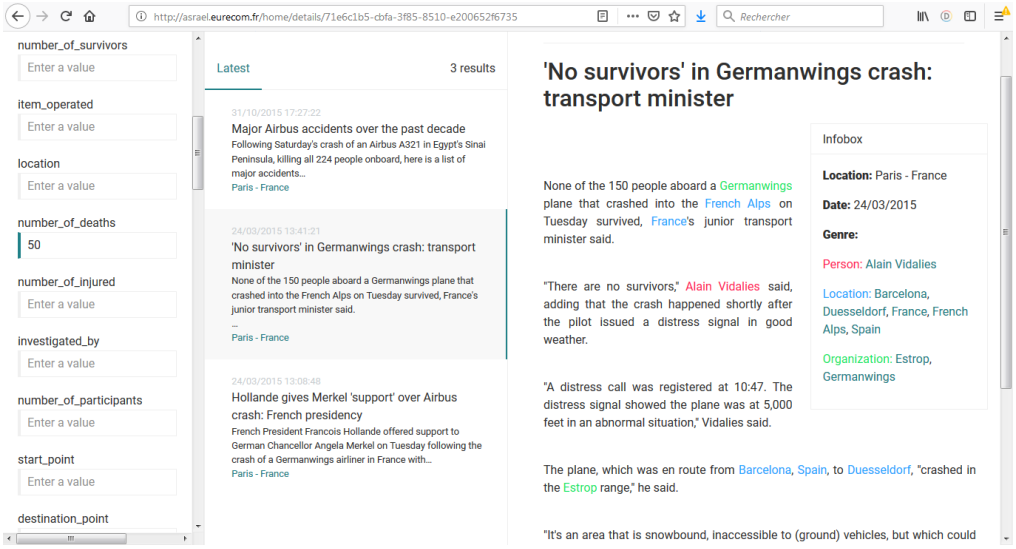

Figure 6: Search engine filtering articles describing plane crash events having caused at least 50 victims.

used as a distant supervision for training an annotation system. This system could then be able to extract the structure of the events from the news article, even if they are not in Wikidata. We also plan to work on a multilingual support for this system.

ACKNOWLEDGEMENTS

This work has been partially supported by the French National Research Agency (ANR) within the ASRAEL Project, under grant number ANR-15-CE23-0018, and the ContentCheck Project, under grant number ANR-15-CE23-0025-01.

REFERENCES

[1] Razvan Bunescu and Marius Paşca. 2006. Using Encyclopedic Knowledge for Named entity Disambiguation. In *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

[2] Simon Gottschalk and Elena Demidova. 2018. EventKG: A Multilingual Event-Centric Temporal Knowledge Graph. In *Extended Semantic Web Conference (ESWC)*.

[3] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. 2013. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. *Artificial Intelligence, special issue on Wikipedia and Semi-Structured Resources* (2013).

[4] Ji Young Lee and Franck Dernoncourt. 2016. Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California, 515–520.

[5] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2013. DBpedia - A Large-scale, Multilingual

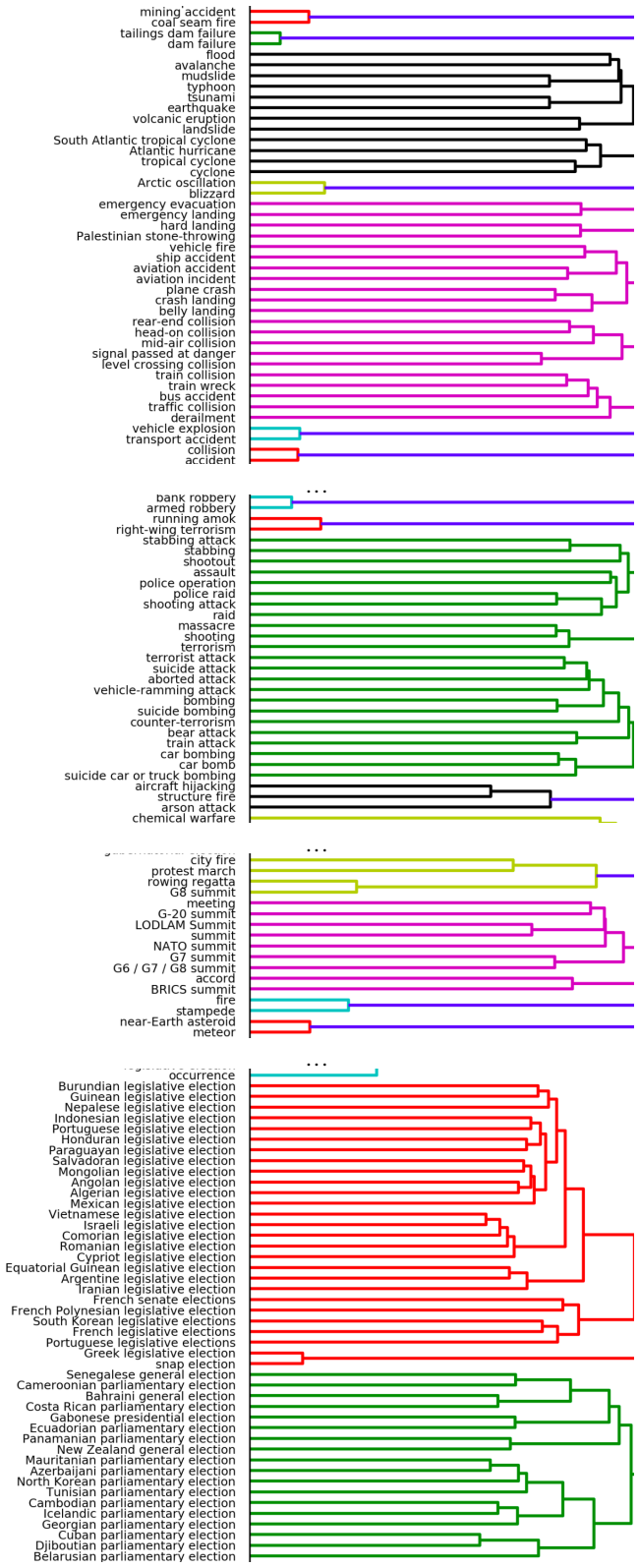

Figure 7: Extracts of the dendrogram issued after clustering.

Knowledge Base Extracted from Wikipedia. *Semantic Web Journal* (2013).

[6] Lijun Lyu and Besnik Fetahu. 2018. Real-time Event-based News Suggestion for Wikipedia Pages from News Streams. In *Wiki Workshop*. Lyon, France, 1793–1799.

[7] Rada Mihalcea and Andras Csomai. 2007. Wikify!: Linking Documents to Encyclopedic Knowledge. In *16th ACM Conference on Conference on Information and Knowledge Management (CIKM)*. Lisbon, Portugal, 233–242.

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems (NIPS)*. 3111–3119.

[9] Arunav Mishra. 2014. Linking Today's Wikipedia and News from the Past. In *7th ACM Workshop on Ph.D Students on Information and Knowledge Management (PIKM)*. Shanghai, China, 1–8.

[10] Arunav Mishra and Klaus Berberich. 2016. Leveraging Semantic Annotations to Link Wikipedia and News Archives. In *38th European Conference on Information Retrieval (ECIR)*. Padua, Italy, 30–42.

[11] Julien Plu, Giuseppe Rizzo, and Raphael Troncy. 2019. ADEL: ADaptable Entity Linking. *Semantic Web Journal* (2019).

[12] Kevin Reschke, Martin Jankowiak, Mihai Surdeanu, Christopher Manning, and Daniel Jurafsky. 2014. Event Extraction Using Distant Supervision. In *9th International Conference on Language Resources and Evaluation (LREC)*. Reykjavik, Iceland.

[13] Mitja Trampuš and Blaz Novak. 2012. The Internals Of An Aggregated Web News Feed. In *Conference on Data Mining and Data Warehouses (SiKDD)*.

[14] Giang Binh Tran and Mohammad Alrifai. 2014. Indexing and Analyzing Wikipedia's Current Events Portal, the Daily News Summaries by the Crowd. In *23rd International Conference on World Wide Web (WWW), Demo Track*. Seoul, Korea, 511–516.

[15] Raphaël Troncy. 2008. Bringing the IPTC News Architecture into the Semantic Web. In *7th International Semantic Web Conference (ISWC)*. Karlsruhe, Germany, 483–498.

[16] Willem Robert van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, and Guus Schreiber. 2011. Design and use of the Simple Event Model (SEM). *Journal of Web Semantics* 9, 2 (2011), 128–136.

[17] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.