Automatic Classification and Relationship Extraction for Multi-Lingual and Multi-Granular Events from Wikipedia

Daniel Hienert[1], Dennis Wegener[1] and Heiko Paulheim[2]

[1] GESIS – Leibniz Institute for the Social Sciences
Unter Sachsenhausen 6-8, 50667 Cologne, Germany
{daniel.hienert, dennis.wegener}@gesis.org

[2] Technische Universität Darmstadt
Knowledge Engineering Group
Hochschulstraße 10, 64283 Darmstadt, Germany
paulheim@ke.tu-darmstadt.de

Abstract. Wikipedia is a rich data source for knowledge from all domains. As part of this knowledge, historical and daily events (news) are collected for different languages on special pages and in event portals. As only a small amount of events is available in structured form in DBpedia, we extract these events with a rule-based approach from Wikipedia pages. In this paper we focus on three aspects: (1) extending our prior method for extracting events for a daily granularity, (2) the automatic classification of events and (3) finding relationships between events. As a result, we have extracted a data set of about 170,000 events covering different languages and granularities. On the basis of one language set, we have automatically built categories for about 70% of the events of another language set. For nearly every event, we have been able to find related events.

Keywords: Historical Events, News, Wikipedia, DBpedia

1 Introduction

Wikipedia is an extensive resource for different types of events like historical events or news that are user-contributed and quality-proven. Although there is plenty of information on historical events in Wikipedia, only a small fraction of these events is available in a structured form in DBpedia. In prior work we have focused on extracting and publishing these events for the use in the semantic web and other applications [6]. In this paper, we focus on how the dataset can be enriched and its quality can be further improved. We address this question with two approaches: to find categories for events and to extract relationships between events. These features can later be used in end-user applications to list related events, browse between events or filter events from the same category.

The remainder of this paper is as follows: Section 2 presents related work. In Section 3, we address the question on how events can be detected, extracted, processed and presented in different forms for the semantic web (Workshop questions 1, 2 and 3). In Section 4 we present an approach on how events can be automatically classified with categories (Question 1). In Section 5 we show how relationships between events from different languages and granularities can be found (Question 1).

2 Related Work

There is a range of systems specialized for the extraction of events and temporal relations from free text. The TARSQI toolkit [16] can detect events, times and their temporal relations by temporal expressions in news articles. HeidelTime [14] is a rule-based system for the extraction and normalization of temporal expressions using mainly regular expressions. The TIE system [9] is an information extraction system that extracts facts from text with as much temporal information as possible and bounding start and end times.

Some work has been done for the extraction of events from Wikipedia articles with machine learning or rule-based approaches and the presentation for the end user in user interfaces with timelines and maps. The approach of Bhole [2] for example first classifies Wikipedia articles as persons, places or organizations on the basis of Support Vector Machines (SVM). Then text mining is used to extract links and event information for these entities. Entities and their events can be shown on a timeline. In another system [3] major events are extracted and classified for a historical Wikipedia article and shown in a user interface with a timeline, map for event locations and named entities for each event.

Other work concentrates on the extension of knowledge bases like DBpedia [1] or YAGO [15] with temporal facts. Exner and Nugues [4] have extracted events based on semantic parsing from Wikipedia text and converted them into the LODE model. They applied their system to 10% of the English Wikipedia and extracted 27,500 events with links to external resources like DBpedia and GeoNames. Since facts in knowledge bases evolve over time the system T-YAGO [17] extends the knowledge base YAGO with temporal facts, so that they can be queried with a SPARQL-style language. As a subsequent technology, Kuzey & Weikum [8] presented a complete information extraction framework on the base of T-YAGO that extracts more than one million temporal facts from Wikipedia resources like semi-structured data (infoboxes, categories, lists and article titles) and free text of Wikipedia articles with a precision over 90% for semi-structured and 70% for full text extraction. Alternatively, the YAGO2 system [7] extends the YAGO knowledge base with temporal and spatial components. This information is extracted from infoboxes and other resources like GeoNames.

There is a collection of ontologies for the modeling of events in RDF like EVENT[1], LODE [13], SEM [5], EventsML[2] and F [12], a comparison can be found in [5].

However, most related work in this field is about the *extraction* of events from free text or knowledge bases like Wikipedia or YAGO and the *enrichment* of entities from text or knowledge bases with temporal information. Not much work has been done on

[1] http://motools.sourceforge.net/event/event.html

[2] http://www.iptc.org/EventsML/

the further enrichment of event datasets such as adding relations or additional information like categorizations.

3 Events from Wikipedia

Wikipedia is a rich data source for events of different topics, languages and granularity. Most research focuses on the extraction of events from the full text of Wikipedia articles and on relating it to the appropriate entities. Major historical events have their own article, or events are collected in articles for a special topic. Events are also collected in time units of different granularity (i.e. years or months) available for different languages. These articles contain lists with events, whose structure is relatively stable. In prior work we have focused on the extraction of events from year-based articles, which include information on individual years for different languages [6]. Table 1 gives an overview over the extracted events for different languages and their extraction quotients. The number of possible events for each language is based on the assumption that every event line in the Wiki markup starts with an enumeration sign. The extracted dataset has several unique characteristics: (1) it has a wide temporal coverage from 300 BC to today, (2) it is available for a lot of different languages, (3) different granularities (year or month) are available, (4) Wikipedia users already have chosen which events are important for different granularities, (5) events already contain links to entities, (6) events have categorizations or can be enriched with categorization and relationships among each other.

Table 1. Number of extracted events for language/granularity and the extraction quotients

Language/Granularity	Possible Events	Extracted Events	Extraction Quotient
German/Year	36,713	36,349	99.01%
English/Year	39,739	34,938	87.92%
Spanish/Year	20,548	19,697	95.86%
Romanian/Year	13,991	10,633	76.00%
Italian/Year	14,513	10,339	71.24%
Portuguese/Year	8,219	7,395	89.97%
Catalan/Year	7,759	6,754	87.05%
Turkish/Year	3,596	3,327	92.52%
Indonesian/Year	2,406	1,963	81.59%
English/Month	38,433	35,633	92.71%
German/Month	11,660	11,474	98.40%
Total		**178,502**	

3.1 Extraction, processing and provision

Figure 1 shows the overall extraction and processing pipeline. Our software crawls Wikipedia articles for different granularities (years and months) and different languages. For year-based articles, German, English, Spanish, Romanian, Italian, Portuguese, Catalan, Turkish and Indonesian with a temporal coverage from 300BC to today are crawled. For daily events, German and English articles from the year 2000 to today are collected. In the next step, the events are extracted from Wiki markup. We use a set of language-dependent regular expressions for the identification of the event section in the article, the identification of events in the event section and the separation of date, description and links for each event. Events can be further described by categories that result from headings in the markup. Events and links are then stored in a MySQL database.

The resulting data set is then further processed. For the automatic classification see Section 4, for the finding of relationships between events see Section 5. We also crawl the Wikipedia API to add an individual image to each event for the use in the timeline.

We provide access to the extracted events via the Web-API, SPARQL endpoint, Linked Data Interface and in a timeline. The Web-API[3] gives lightweight and fast access to the events. Events can be queried by several URL parameters like begin_date, end_date, lang, query, format, html, links, limit, order, category, granularity and related. Users can query for keywords or time periods, and results are returned in XML or JSON format. The Linked Data Interface[4] holds a representation of the yearly English dataset in the LODE ontology [13]. Each event contains links to DBpedia entities. Users can query the dataset via the SPARQL endpoint (http://lod.gesis.org/historicalevents/sparql). Additionally, yearly events for the English, German and Italian dataset are shown in a Flash timeline (http://www.vizgr.org/historical-events/timeline/) with added images and links to Wikipedia articles. Users can search for years, scroll and scan the events and navigate to Wikipedia articles.

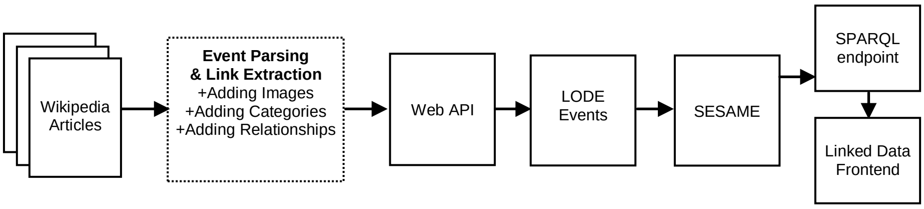

Fig. 1. Processing, extraction and provision pipeline.

3.2 Extraction of daily events

In addition to the extraction of yearly events presented in [6], we have extracted daily events from the German and English Wikipedia version. The German version provides events on a daily basis in articles of months (i.e. http://de.wikipedia.org/wiki/Juni_2011) from the year 2000 to today. The English structure is quite more complicated and daily events are distributed in three different site structures: (1) most daily events are collected in the *Portal:Current events* (http://en.wikipedia.org/wiki/Portal:Current_events), (2) some events are collected in the *Portal:Events* (before July 2006) and (3) other events are collected in month collections similar to the German version. English daily events are also available for the years 2000 to today. First, we have extended the extraction software to query

[3] http://www.vizgr.org/historical-events/

[4] http://lod.gesis.org/pubby/page/historicalevents/

these site structures. Then, regular expressions for the identification of event section and for the individual events have been added. The extraction algorithm had to be slightly modified to handle new structures specific for daily events. As a result, the software could extract 35,633 English daily events (extraction quotient: 92.17%) and 11,747 German daily events (extraction quotient: 98.40%).

3.3 Analyzing the data set

The overall data set has been analyzed as a prerequisite to the automatic classification and the search for relationships between events. The number of extracted events and extraction quotients for different languages and granularity are shown in Table 1. The categories in German events are created from subheadings on the corresponding Wikipedia page. Yearly German events are categorized with one or two categories by headings of rank 2 or 3, which can be used for the automatic classification of events. Table 2 shows the ten most used categorizations for German events. In English or other languages categorizations are rarely used. The number of links and entities per event can be seen in Table 3. In the German and English dataset most events have between one and four links.

Table 2. Categories (translated) and their counts for yearly German events

Category	Count
Politics and world events	18,887
Culture	4,135
Science and technology	3,096
Religion	2,180
Economy	2,011
Sports	1,434
Disasters	1,351
Politics	613
Culture and Society	309
Society	286

Table 3. Distribution of links to entities within the German and English yearly dataset

Count of entities	English	German
No entity	6,371	1,489
One entity	5,773	7,815
Two entities	10,143	9,969
Three entities	8,405	8,086
Four entities	4,499	4,606
Five entities	2,376	2,457
Six entities	1,271	1,234
Seven or more entities	901	693

4 Automatic Classification of Events

To provide a useful semantic description of events, it is necessary to have types attached to these events. Possible types could be "Political Event", "Sports Event", etc. In the crawled datasets, some events already have types extracted from the Wikipedia pages, while others do not. Therefore, we use machine learning to add the types where they are not present.

The datasets we have crawled already contain links to Wikipedia articles. In order to generate useful machine learning features, we have transformed these links to DBpedia entities. For inferring event types, we have enhanced our datasets consisting of events and their descriptions by more features: the direct types (rdf:type) and the categories (dcterms:subject) of the entities linked to an event, both including their transitive closures (regarding rdfs:subClassOf and skos:broader, respectively). For enhancing the datasets, we have used our framework FeGeLOD [11], which adds such machine learning features from Linked Open Data to datasets in an automatic fashion. The rationale of adding those features is that the type of an event can be inferred from the types of the entities involved in the event. For example, if an entity of type SoccerPlayer is involved in an event, it is likely that the event is a sports event.

As discussed above, the majority of events in our datasets comprises between one and four links to entities. Therefore, we have concentrated on such events in our analysis. We have conducted two experiments: first, we have inferred the event types on events from the German dataset, using cross validation for evaluation. Second, we have learned models on the German datasets and used these models to classify events from the English dataset, where types are not present. In the second experiment, we have evaluated the results manually on random subsets of the English dataset.

Figure 2 depicts the classification accuracy achieved in the first experiment, using 10-fold cross validation on the German dataset. We have used four random subsets of 1,000 events which we have processed by adding features and classifying them with three different commonly used machine learning algorithms: i.e., Naïve Bayes, Ripper (in the JRip implementation), and Support Vector Machines (using the Weka SMO implementation, treating the multi-class problem by using 1 vs. 1 classification with voting). As a baseline, we have predicted the largest class of the sample. It can be observed that the categories of related entities are more discriminative than the direct types. The best results (around 80% accuracy) are achieved with Support Vector Machines.

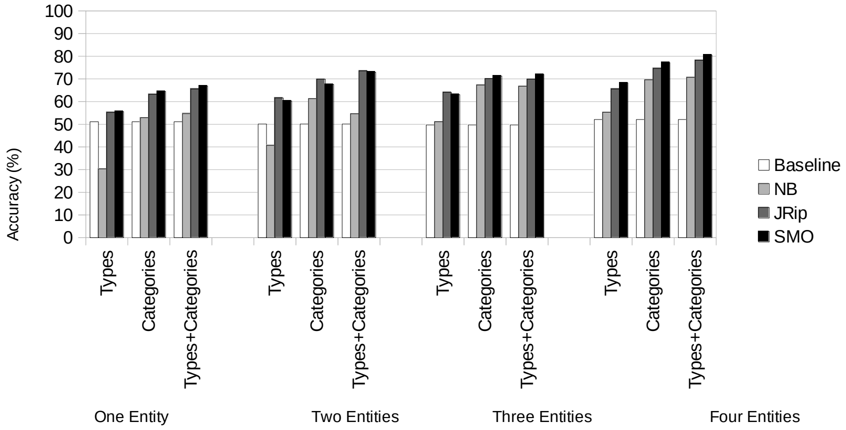

Fig. 2. Classification accuracy on the German dataset, using ten-fold cross validation for evaluation

Since Support Vector Machines have yielded the best results in the first experiment, we have trained four SVMs for the second experiment, one for each number of related entities (one through four), using the subsets of 1,000 events. We have then used these models to classify four subsets of the English dataset, consisting of 50 events each. The results of that classification have been evaluated manually.

The results are shown in Figure 3. First, we have tested the best performing combination of the first experiment, using both categories and direct types of the related entities. Since the results were not satisfying, we have conducted a second evaluation using only direct types, which yielded better results. The most likely reason why categories work less well as features than classes is that the German and the English DBpedia use the same set of classes (i.e., DBpedia and YAGO ontology classes, among others), but different categories. In our experiments, we have observed that only a subset of the categories used in the German DBpedia have a corresponding category in the English DBpedia. Thus, categories, despite their discriminative power in a single-language scenario, are less suitable for training cross-language models.

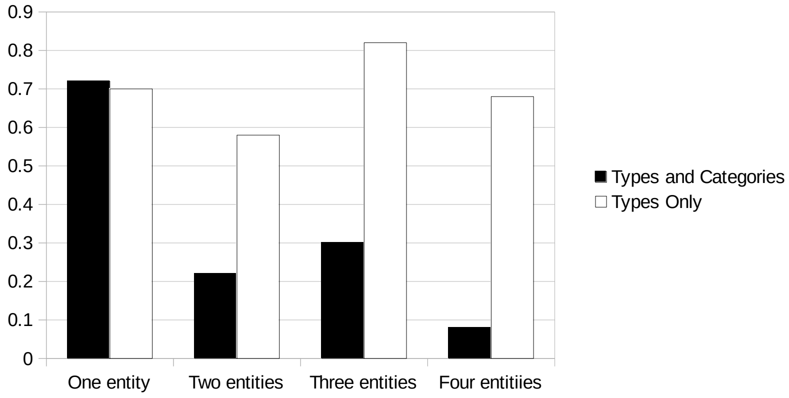

Fig. 3. Classification accuracy achieved on English dataset, using Support Vector Machines trained on the German dataset

In summary, we have been able to achieve a classification accuracy of around 70% for the English dataset, using a model trained on the German dataset. The results of both experiments show that machine learning with features from DBpedia is a feasible way to achieve an automatic classification of the extracted events.

5 Relationships Between Events

With a dataset of events for different languages and granularities it is interesting to know which relations between these events exist. To find relationships, different features of the events could be used: (1) time, (2) categories, (3) topic/content or (4) links. Time as a single criterion is not by far enough. The category is too simplistic and there are only a few categories. Relationships based on the topic/content of the event are not easy to find as the events only include micro-text with a few words or sentences. Taking links as a criterion, we have to consider which links to take and how many links. In our approach we use a combination of the features time and links for extracting relationships between events.

As described in Section 3.3, we have extracted 178,502 events in total. From these, 172,189 events include links. As a preprocessing step, we transform every non-English link to the English equivalent by querying the inter-language link from the Wikipedia API. As a result, every event from different languages contains links to English Wikipedia/DBpedia entities.

In the following, we analyze this set of events. As first step we vary the number of links that two events have to share and count the events that share this number of links with at least one other event (see Table 4). In detail, we consider two events to *share a link* if these events contain a link to the same DBpedia entity. From our analysis results it can be seen that 95.8 % of the events (that include links) share at least one link with at least one other event. As we are dealing with a multi-lingual set of events, it is interesting to know how many events share one link with at least one event of a different language. In our set of events, 155,769 events share at least one link with at least one other event of a different language, which is 90.5 % of the events in the set. 75.7% of the events include a link to another granularity, i.e. from year to month or vice versa.

Table 4. Analysis of the number of shared links between events

# shared links	# events that share the number of links with at least one other event	in % (# total events = 172,189)
1	165,014	95.8 %
2	100,401	58.3 %
3	35,456	20.6%
4	9,900	5.7%

So far, we have looked for events that share one link in the overall database. In the following, we vary the time interval in which we search for these events (see Table 5). In detail, if we look at an event at time x, an interval of one month means that we search for events in the time interval [x-15 days : x + 15 days]. For the time-based analysis, we can only consider events where the date includes information on the day (and not only on the month and year). In our set these are 109,510 events.

Table 5. Analysis of the number of events that hold shared links in a given time interval

Time interval	Number of events that share one link with at least one other event in the time interval	In % (number of total events with exact date = 109,510)
Overall	105,042	95,9 %
Year [x-182 days : x+182 days]	90,193	82,4 %
Month [x-15 days : x+15 days]	74,499	68,0 %
Week [x-3 days : x+3 days]	61,246	55,9 %

Based on this analysis we have been able to define the *relatedness* between two events A and B with the time interval minimal and the number of shared links maximal between these events. Whereby we have found that in our dataset, a large part has at least one link in common (95.8%) within a time interval of a year (82.4%) and we can also find links to other languages (90.5%) and granularities (75.7%).We have implemented the relatedness feature in the Web-API. To compute related events for an individual event, we query for events that have at least one link in common within a time interval of plus/minus ten years and then sort results first by number of shared links and then by time distance to the original event.

For example, the query for *Arab Spring*[5] finds eleven events from the yearly English dataset and related events from other languages and granularities. For example, the event of 2011/01/14: “Arab Spring: The Tunisian government falls after a month of increasingly violent protests President Zine El Abidine Ben Ali flees to Saudi Arabia after 23 years in power.” lists equivalent events from different languages, i.e. Italian: “In Tunisia, dopo violente proteste…”, Spanish: “en Túnez el presidente Zine El Abidine Ben…”, German: “Tunis/Tunesien: Nach den schweren Unruhen der Vortage verhängt Präsident Zine el-Abidine…” and from a month/news view: “Thousands of people protest across the country demanding the resignation of President Zine El Abidine Ben Ali. [Link] (BBC)”

As a final step we have compiled an evaluation set with 100 events and 5 related events for each and analyzed them manually. We have found that the perceived relatedness between two events (1) depends on the time interval between events and (2) depends on the count (1 vs. 4), type (general types like *Consul* vs. finer types like *Julius Caesar*) and position (at the beginning or the end of the description) of shared links.

In summary, we have been able to find a related event for nearly every event in the dataset, also for events from other languages and granularities.

6 Conclusion

We have extracted an event dataset from Wikipedia with about 170,000 events for different languages and granularities. A part of these events includes categories which can be used to automatically build categories for about 70% of another language set on the basis of links to other Wikipedia/DBpedia entities. The same linking base is used together with a time interval to extract related events for nearly every event, also for different languages and granularities.

At the moment, we only use Wikipedia/DBpedia links that are already included in the events' descriptive texts. However, those links are not always complete or available in other data sets. Using automatic tools such as DBpedia spotlight [10] would help increasing the result quality and allow us to process text fragments without hyperlinks as well.

At the end of Section 5 we have shown that the perceived quality of events depends also on the abstractness of links. The analysis on how the abstractness of links can be modeled and used as an additional feature for the ranking of related events remains to future work.

[5] http://www.vizgr.org/historical-events/search.php?query=arab%20spring&related=true

References

1. Auer, S. et al.: DBpedia: A Nucleus for a Web of Open Data. In 6th Int'l Semantic Web Conference, Busan, Korea. pp. 11–15 Springer (2007).
2. Bhole, A. et al.: Extracting Named Entities and Relating Them over Time Based on Wikipedia. Informatica (Slovenia). 31, 4, 463–468 (2007).
3. Chasin, R.: Event and Temporal Information Extraction towards Timelines of Wikipedia Articles. Simile. 1–9 (2010).
4. Exner, P., Nugues, P.: Using semantic role labeling to extract events from Wikipedia. Proceedings of the Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011). Workshop in conjunction with the 10th International Semantic Web Conference 2011 (ISWC 2011). , Bonn (2011).
5. Hage, W.R. van et al.: Design and use of the Simple Event Model (SEM). Web Semantics: Science, Services and Agents on the World Wide Web. 9, 2, (2011).
6. Hienert, D., Luciano, F.: Extraction of Historical Events from Wikipedia. Proceedings of the First International Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data (KNOW@LOD 2012). , Heraklion, Greece (2012).
7. Hoffart, J. et al.: YAGO2: exploring and querying world knowledge in time, space, context, and many languages. Proceedings of the 20th international conference companion on World wide web. pp. 229–232 ACM, New York, NY, USA (2011).
8. Kuzey, E., Weikum, G.: Extraction of temporal facts and events from Wikipedia. Proceedings of the 2nd Temporal Web Analytics Workshop. pp. 25–32 ACM, New York, NY, USA (2012).
9. Ling, X., Weld, D.S.: Temporal Information Extraction. In: Fox, M. and Poole, D. (eds.) AAAI. AAAI Press (2010).
10. Mendes, P. et al.: DBpedia Spotlight: Shedding Light on the Web of Documents. In the Proceedings of the 7th International Conference on Semantic Systems (I-Semantics). (2011).
11. Paulheim, H., Fürnkranz, J.: Unsupervised Generation of Data Mining Features from Linked Open Data. International Conference on Web Intelligence and Semantics (WIMS'12). (2012).
12. Scherp, A. et al.: F–a model of events based on the foundational ontology dolce+DnS ultralight. Proceedings of the fifth international conference on Knowledge capture. pp. 137–144 ACM, New York, NY, USA (2009).
13. Shaw, R. et al.: LODE: Linking Open Descriptions of Events. Proceedings of the 4th Asian Conference on The Semantic Web. pp. 153–167 Springer-Verlag, Berlin, Heidelberg (2009).
14. Strötgen, J., Gertz, M.: HeidelTime: High quality rule-based extraction and normalization of temporal expressions. Proceedings of the 5th International Workshop on Semantic Evaluation. pp. 321–324 Association for Computational Linguistics, Stroudsburg, PA, USA (2010).
15. Suchanek, F.M. et al.: Yago: a core of semantic knowledge. Proceedings of the 16th international conference on World Wide Web. pp. 697–706 ACM, New York, NY, USA (2007).
16. Verhagen, M., Pustejovsky, J.: Temporal processing with the TARSQI toolkit. 22nd International Conference on on Computational Linguistics: Demonstration Papers. pp. 189–192 Association for Computational Linguistics, Stroudsburg, PA, USA (2008).
17. Wang, Y. et al.: Timely YAGO: Harvesting, Querying, and Visualizing Temporal Knowledge from Wikipedia. Proceedings of the 13th International Conference on Extending Database Technology (EDBT), Lausanne, Switzerland, March 22-26. pp. 697–700 (2010).