Towards a Complete Event Type Taxonomy

Aljaž Košmerlj
Jožef Stefan Institute
Jamova 39
Ljubljana, Slovenia
aljaz.kosmerlj@ijs.si

Evgenia Belyaeva
Jožef Stefan Institute
Jamova 39
Ljubljana, Slovenia
jenya.belyaeva@ijs.si

Gregor Leban
Jožef Stefan Institute
Jamova 39
Ljubljana, Slovenia
gregor.leban@ijs.si

Blaž Fortuna
Ghent University - iMinds
Ghent, Belgium
blaz.fortuna@intec.ugent.be

Marko Grobelnik
Jožef Stefan Institute
Jamova 39
Ljubljana, Slovenia
marko.grobelnik@ijs.si

ABSTRACT

We present initial results of our effort to build an extensive and complete taxonomy of events described in news articles. By crawling Wikipedia's current events portal we identified nine top–level event types. Using articles referenced by the portal we built a event type classification model for news articles using lexical and semantic features and present a small–scale manual evaluation of its results. Results show that our model can accurately distinguish between event types but its coverage could still be significantly improved.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*

Keywords

natural language processing; event extraction; event type; news classification; Wikipedia

1. INTRODUCTION

The volume of news has long since surpassed human ability to consume it in its entirety and recently we have seen a rise of automatic tools and services that attempt to organize or recommend news for the reader. Within our group we have developed the Event Registry[1] [11], which is able to identify sets of news articles describing same real–world events. Important entities and keywords are also identified but we do not know the roles they play in the events. For example we may know that the *German football team* was an important entity for the *2014 FIFA World Cup finals* but we do not know that it was the winning team.

In order to extract that kind of information we need a schema for describing events and a way of populating it. Unfortunately the performance of automatic event schema generation and population approaches [3, 2, 4] is still not good enough for practical use, and no complete, manually curated event type schema exists according to our knowledge. Event schemas do exist[13] but they are too abstract for our use. This paper presents the initial results of our attempt to build an event type taxonomy in a top–down manner. We are using Wikipedia's current events portal as a data source. The portal contains a chronological list of events with links to source articles organised into several user–defined event types. We have manually cleaned the event type set and crawled the source links to obtain a dataset of news articles with labeled event types. We validate this data by building a classification model and testing it using data from a comprehensive crawl of news articles covering four months.

The rest of the paper is organized as follows. First we continue with a brief discussion of related work in subsection 1.1. The data we used in our experiments along with experimental methodology and results are described in section 2. Finally we conclude and outline future work in section 3.

1.1 Related work

As already mentioned there exist event description schemas (a nice list of them and their comparison can be found in [13]), but they are too high–level for our purposes. Our aim is a complete taxonomy of event types and their attributes like specific roles entities play in the events etc. Think *football match* and *the winning team* or *judging referee* as opposed to *temporal_event* and *involved_agent*.

Wikipedia has already been used as a data source for extracting event information, but most approaches use Wikipedia article content [13, 9, 6]. Similar approaches aim towards population of knowledge bases like YAGO with canonicalized event information by linking events from news to Wikipedia event categories which are then mapped to WordNet classes [8, 10]. All these are limited in coverage by the availability of Wikipedia articles about specific events and the coverage of vocabularies like WordNet. There is also a difference in language properties of news article text and the encyclopedic text of Wikipedia.

[1]http://eventregistry.org/

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742005.

2. EXPERIMENTS AND RESULTS

2.1 Data

Wikipedia, the online encyclopedia, contains a *current events* portal[2] with a chronological list of events that spans back to the beginning of 1998. The list contains short descriptions of events which are due to the community–driven nature of the project of varying quality but since 2005 typically contain a reference to a news item describing the event. Since mid 2010 each event is designated a high–level type depending on the topic (e.g. sport, science etc.). These event types are user–defined and are not curated. By crawling the entire event archive in the time span from August 2010 to August 2014 we collected all event types and manually identified nine main event types listed in Table 1. These main event types were obtained by merging some event types (e.g. merging *attacks* into *armed conflicts and attacks* which had significantly more events) and removing those with a small number of events.

event type	*nr. of articles*
armed conflicts and attacks	3 516
arts and culture	744
business and economy	963
disasters and accidents	1 851
health and environment	115
law and crime	1 907
politics and elections	3 180
science and technology	521
sport	1 086
total	13 883

Table 1: Event types with numbers of their articles.

We also crawled the reference links in the event items downloading the referenced news articles (skipping links to other media types like YouTube videos or pdf documents), removing the HTML chrome and putting them through a lexical and semantic analysis pipeline [1]. Table 1 contains the final number of fully annotated articles of each event type. We use these articles as a learning dataset to build a model which classifies news articles into event types.

As a test–bed we use articles from a newsfeed service[3] [14] which collects news articles from a large number of RSS feeds and processes them with the same lexical and semantic analysis pipeline as mentioned in the previous paragraph. In our experiment we use a dataset of 8 168 745 articles from a four month period from January to April 2014.

2.2 Event type classification

Matching event types with articles was done using standard text classification methodology. We used vector space model [12] to represent news articles, with TFIDF unigram and bigram features, Porter stemmer and a standard list of stop words. We also tried additional representation, where the word vocabulary is extended with meta–data available for all articles: entities, categories and tags. Note that all meta–data fields were added to the articles automatically by a named entity recognizer and by an automatic classifier into DMoz taxonomy.

[2]https://en.wikipedia.org/wiki/Portal:Current_events
[3]http://newsfeed.ijs.si/

Support Vector Machines (SVM) [5] with linear kernel was used as learning method. We tried several options for the value of SVM cost parameter C and report on results for 1 and 10. Each event type had an associated binary classification model, which was trained using one–vs–all approach: articles with the event type were used as positive examples, and articles from all other event types were used as negative examples.

We performed a 10–fold cross validation in order to evaluate trained models. First, each article was randomly assigned to one of ten bins. Second, in turn we head out each bin, trained classification models trained on the remaining nine bins, and evaluated the models on the held–out bin.

Table 2 shows the results of cross validation for two values of SVM cost parameter C and for two version of features space: with and without meta–data. The results are macro averages over event types. It can be seen that higher cost parameter improves the precision, but the F_1 score is overall higher in the case of lower cost parameter. Meta–data features were found to degrade performance in this setting. Table 3 shows the performance of best performing classifier ($C = 1$ and no meta–data features) for individual event types. It can be seen that articles about sports and accidents are the easiest to identify, whereas articles about environment are the hardest to identify.

features	C	*precision*	*recall*	F_1
text	1	0.82	0.75	0.77
text	10	0.86	0.68	0.75
text + meta–data	1	0.82	0.67	0.73
text + meta–data	10	0.79	0.66	0.72

Table 2: Results for all features, $j = 10$.

category	*precision*	*recall*	F_1
armed conflicts and attacks	0.79	**0.91**	**0.85**
arts and culture	0.82	0.59	**0.69**
business and economy	0.78	**0.72**	**0.75**
disasters and accidents	0.93	**0.91**	**0.92**
health and environment	**0.84**	0.37	0.52
law and crime	0.70	**0.78**	**0.74**
politics and elections	0.69	**0.87**	**0.77**
science and technology	**0.91**	0.71	**0.80**
sport	**0.94**	**0.91**	**0.93**
average	0.82	**0.75**	**0.77**

Table 3: Results for text features, $C = 1$ and $j = 10$.

2.3 Manual validation

To validate the model on a realistic news dataset we applied the best performing classifier to the newsfeed dataset of 8 million articles described in Section 2.1. Each article from the dataset was classified to zero or more event types. Table 4 shows the number of articles classified into each of the event types.

In order to get an estimate of the quality of the model we performed two types of manual validation. Because of the large number of class values (i.e. event types), the size of the test set and the time–consuming nature of manual validation (each article has to be read by a human reader) we did not perform a general estimation of accuracy by using

event type	*nr. of articles*
sport	1 295 698
business and economy	965 805
politics and elections	945 165
law and crime	877 323
arts and culture	601 111
disasters and accidents	396 047
armed conflicts and attacks	303 122
science and technology	121 397
health and environment	12 785
none	2 965 199

Table 4: Numbers of articles classified with individual event types.

a random sample of all articles but performed the validation with more targeted samples.

To see how well the articles fit into their classified event types we randomly selected ten articles that received a positive score from the classifier for each event type resulting in a validation set of 90 articles. A human then read all the articles and decided if the event type, the article was classified in, was appropriate (1) or not (0). Event types of only four articles were deemed inappropriate by the human reader meaning 96% of articles from the sample were classified in the correct event type. Of the four wrongly classified articles two were classified in *politics and elections* event type and one in *armed conflicts and attacks* and *law and crime* each. The article wrongly classified into *armed conflicts and attacks* belonged into *politics and elections* whereas the rest described events of more general societal nature with topics like religion and education.

The purpose of the second round of manual evaluation was to estimate the coverage of the model. That is, how many articles not classified by the model actually do not describe events. We randomly sampled 100 articles that received a negative score by the model and did not fit into any event type according to the model. A human reader then read all the articles and decided whether:

- the article belonged to one of the event types in our model,
- did not belong to one of the event types in our model but still described an event,
- did not describe an event.

The results are listed in table 5. The results show that 63% of the articles should be classified in one of the event types. Most of the rest of the articles are not events with only 7% of the articles describing events of a type not present in the model. The articles describing events of types missing in the model are again addressing societal topics like education and gossip.

3. CONCLUSIONS

Using data collected by crawling Wikipedia's current events portal we were able to identify nine main event types for events described in news articles. We built and tested several models that classified articles into event types using text and meta–data features. Results have shown that meta–data features confuse the model and decrease performance. Manual evaluation indicates that our model has high precision

event type	*nr. of articles*
armed conflicts and attacks	2
arts and culture	16
business and economy	10
disasters and accidents	5
health and environment	10
law and crime	5
politics and elections	3
science and technology	9
sport	3
other event type	7
not an event	30

Table 5: Numbers of articles of particular event type as determined by the second round of manual evaluation

but its recall could still be significantly improved. Several articles which described events and did not fit into any event type in our model were found during the manual evaluation. These events are mostly of a societal nature (e.g. religious observances, "gossip column" events etc.) which appear to be under–represented in the Wikipedia's current events portal. This indicates the direction of further extensions of our event types taxonomy.

3.1 Future work

Most obvious future tasks is are improving recall of the model and extending the event types with those of more societal nature. The long–term aim of our work is construction of an extensive and complete event taxonomy where each event is described with a schema detailing roles of entities in that event. The classification presented in this paper represents the top–level split of the taxonomy into very general event types. Deeper, more detailed levels will be partially built using combined knowledge of existing knowledge bases (OpenCyc, Framenet etc.). In order to fill the blind spots not covered by existing ontologies, especially in the long tail of event types, we are developing a semi–automatic crowdsourcing interface [7] capable of building event schema and extracting event information. Event type classification models like the one described in this paper will be used to guide users of the interface by producing recommendations for schema extensions and document annotation.

4. ACKNOWLEDGMENTS

This work was funded by the European Union through project XLike (FP7-ICT-2011-288342).

5. REFERENCES

[1] X. Carreras, L. Padró, L. Zhang, A. Rettinger, Z. Li, E. García-Cuesta, v. Agić, B. Bekavec, B. Fortuna, and T. Štajner. Xlike project language analysis services. In *Proceedings of the Demonstrations Session at EACL 2014*, pages 9–12, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

[2] N. Chambers. Event schema induction with a probabilistic entity-driven model. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, 2013.

[3] N. Chambers and D. Jurafsky. Template-Based Information Extraction without the Templates.

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 976–986, 2011.

[4] J. C. K. Cheung, H. Poon, and L. Vanderwende. Probabilistic Frame Induction. *Proceedings of 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 837–846, 2013.

[5] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*. Cambridge University Press, New York, NY, USA, 2000.

[6] P. Exner and P. Nugues. Using semantic role labeling to extract events from wikipedia. In *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*, pages 23–24, 2011.

[7] A. Košmerlj, J. Belyaeva, G. Leban, B. Fortuna, and M. Grobelnik. Crowdsourcing event extraction. In *NewsKDD: Data Science for News Publishing workshop. Workshop in conjunction with KDD2014 the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2014.

[8] E. Kuzey, J. Vreeken, and G. Weikum. A fresh look on knowledge bases: Distilling named events from news. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1689–1698, New York, NY, USA, 2014. ACM.

[9] E. Kuzey and G. Weikum. Extraction of temporal facts and events from wikipedia. In *Proceedings of the 2Nd Temporal Web Analytics Workshop*, TempWeb '12, pages 25–32, New York, NY, USA, 2012. ACM.

[10] E. Kuzey and G. Weikum. Evin: Building a knowledge base of events. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 103–106, Republic and Canton of Geneva, Switzerland, 2014. International World Wide Web Conferences Steering Committee.

[11] G. Leban, B. Fortuna, J. Brank, and M. Grobelnik. Event registry: Learning about world events from news. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 107–110. International World Wide Web Conferences Steering Committee, 2014.

[12] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999.

[13] R. Shaw, R. Troncy, and L. Hardman. Lode: Linking open descriptions of events. In A. Gómez-Pérez, Y. Yu, and Y. Ding, editors, *The Semantic Web*, Lecture Notes in Computer Science, pages 153–167. Springer Berlin Heidelberg, 2009.

[14] M. Trampuš and B. Novak. The internals of an aggregated web news feed. In *Proceedings of 15th Multiconference on Information Society 2012 (IS-2012)*, Ljubljana, Slovenia, 2012.