Classifying Short Descriptions of Past Events

Yasunobu Sumikawa[1(✉)] and Adam Jatowt[2]

[1] Department of Information Sciences, Tokyo University of Science, Tokyo, Japan
ysumikawa@acm.org
[2] Department of Social Informatics, Kyoto University, Kyoto, Japan
adam@dl.kuis.kyoto-u.ac.jp

Abstract. Mentions and brief descriptions of events often appear in a variety of document genres such as news articles containing references to related events, historical accounts or biographies. While event categorization has been previously studied, it was usually done on entire news articles or longer event descriptions. In this work we focus on short descriptions of historical events which are typically in the form of one or a few sentences. We categorize them into 9 general event categories using a range of diverse features and report F-measure close to 80%.

Keywords: Event classification · Short event descriptions
Digital history

1 Introduction

Many past events are referred to in texts in the form of brief references, typically, a sentence or few sentences long. For example, a news article on a recent earthquake can briefly refer to a past earthquake to provide necessary background information. A document about the history of a city would typically mention several past events that affected or occurred in that city. Note that brief descriptions of retrospective events do not need to occur within longer texts, but they may be standalone such as in the timelines or event lists. For example, the Wikipedia's Current Portal[1] contains lists of significant events in each month where every event is usually described by a single brief list item. Table 1 shows examples of event descriptions in the Wikipedia's Current Portal.

We focus in this work on the problem of categorizing short descriptions of important, retrospective events. Correctly understanding event mentions could have many applications. For example, by being able to tell the category of mentioned events one could better understand as well as represent the intricate network of related events thanks to studying which past event types are mentioned in news articles. Furthermore, the lists of historical events or timelines could be structured by organizing the events based on their semantic categories.

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

© Springer International Publishing AG, part of Springer Nature 2018
G. Pasi et al. (Eds.): ECIR 2018, LNCS 10772, pp. 729–736, 2018.
https://doi.org/10.1007/978-3-319-76941-7_69

Table 1. Average lengths, sizes and examples of event descriptions for all the categories. The abbreviated names of classes are used: `Armed Conflicts` and `Attacks` (`AA`), `Arts` and `Culture` (`AC`), `Business` and `Economy` (`BE`), `Disasters` and `Accidents` (`DA`), `Health` and `Environment` (`HE`), `Law` and `Crime` (`LC`), `Politics` and `Elections` (`PE`), `Science` and `Technology` (`ST`) and `Sport` (`S`)

Class	Ave. len.	Num. of events	Example
AA	23.6	8,886	Bombs across Iraq detonate, killing 18 people
AC	22.9	1,800	The Beatles release their back catalogue on iTunes
BE	23.6	2,517	Brazil's economy falls into recession
DA	23.1	4,961	A bus crashes into a ravine in Tibet, killing at least 44 people
HE	28.7	487	The number of Zika virus infected in Singapore rises above 40
LC	27.5	4,984	The Constitutional Council of France upholds a ban on fracking
PE	25.2	5,517	Voters in Costa Rica go to the polls for a general election
ST	24.6	1,066	Iran successfully puts the Fajr satellite in orbit using a Safir-B1 rocket
S	23.3	2,400	The Winter Olympics in Sochi, Russia officially concludes

Equipped with knowledge on the categories of past event mentions one could also foster collective memory studies [1] as well as support search methods for finding historical events. Finally, the classification technique could be used for constructing thematic timelines or event lists (e.g., list of disasters/accidents in Asia, timeline of armed conflicts in USA).

Note that the task is not trivial. Prior literature on event classification typically focused on entire news articles which usually contain sufficient amount of text for effective category assignment [4]. In our case, events can be just passing mentions or can be only briefly described with little content available for their automatic classification. The main challenge lies in the scarcity of data, the ambiguity of expressions and variety of diverse means in which events can be referred to. Furthermore, oftentimes, in realistic scenarios, events are not called by their explicit names, or, they may have no known names[2]. Consequently, their automatic detection using NER tools is problematic. To provide sufficient data we use a range of features including ones computed from external knowledge bases like Wikipedia and VerbNet, ones based on lexical analysis as well as ones based on distributional word representation using neural networks. We make an assumption that the context of such descriptions (e.g., surrounding sentences in

[2] Usually, only very popular or important events have own names.

original text) is not available to cover also the case of standalone descriptions. Hence we rely only on the event description itself.

In prior literature, two kinds of approaches for short text classifiers were assumed. On one hand, context information was added to feature vectors built from text content. For example, Sriram *et al.*'s [10] approach classifies tweets by using author information, url and hashtags of tweets. Nie *et al.* [8] use Naive Bayes classifier equipped with texts, image and video contents for Q&A classification. Lee *et al.* [6] classifies queries using user-click behavior to identify user goals in web search. On the other hand, several studies have used external knowledge bases such as Wikipedia. Zelikovitz and Marquez [12] trained a classifier with LSA based on Wikipedia data, and Phan *et al.* [9] proposed a generalized framework of classifiers with topic model. This framework first trains the topic model on texts of an external resource. Explicit Semantic Analysis (ESA) was applied in [11] to map short texts to Wikipedia articles.

In our study, we classify brief mentions of past events using Wikipedia and other external resources, and we investigate which features (e.g., entities, actions, etc.) are best suited for this task. We test the proposed categorization method on the set of 32,362 short descriptions of events from the last 6 years, where descriptions contain on average 25 words. Our approach achieves on average 79.7% F-measure value which should be sufficient for some applications.

2 Data Collection

Event Classes. We use 9 general event classes[3]: `Armed Conflicts` and `Attacks` (`AA`), `Arts` and `Culture` (`AC`), `Business` and `Economy` (`BE`), `Disasters` and `Accidents` (`DA`), `Health` and `Environment` (`HE`), `Law` and `Crime` (`LC`), `Politics` and `Elections` (`PE`), `Science` and `Technology` (`ST`) and `Sport` (`S`). They were described in [4] as a proposal of a comprehensive event class list based on definitions and guidelines used by Wikipedia editors. Although the authors investigated also automatic classification, they did it on entire news articles and using only simple features (TF-IDF).

Dataset. We collected 32,618 typed events from the *Current Events* portal of Wikipedia. Their timespan ranges from 2010/1/1 to 2016/12/17. The average lengths of event descriptions per individual category are shown in Table 1. On average, for all the classes, the descriptions contain 25 words, though the length can be as short as 10 words. Note that the *Current Events* portal of Wikipedia contains also quite large number (precisely, 69,554) of unlabeled events which we did not collect. They occurred between AD1 and Dec. 2016. Their automatic labeling (or at least automatic support of manual annotation) would be an opportunity for the application of the developed classifiers.

[3] See Table 1 for examples of events in each class.

3 Methodology

In this section we list the features used for constructing the classifiers and give the intuition behind their choice.

Term based features. We first create TF-IDF vectors (F_1) from all the event descriptions to consider representative terms of events.

Latent semantic based features. To capture latent semantic structures present in text we use both Doc2Vec [5] (F_2) and LSA [2] (F_3).

Verbs based features. Verbs are the essence of events indicating what actions were carried. We then map verbs to VerbNet[4] to obtain their semantic classes (e.g., `destroy` class contains `demolish`, `ruin` verbs among others) and count the number of collected classes to use as features (F_4). There are 429 semantic groups of verbs allowing to organize events by their common actions. We use Stanford POS tagger to collect verbs.

Entity type based features. Most event descriptions contain entities (e.g., persons, organizations, or places), which are actors, locations where the events occurred, important stakeholders and so on. Certain entity types can be strongly associated with particular event classes (e.g., the occurrence of company or organization may suggest `Business & Economy` type of an event). Similarly, the lack of a particular entity type can be suggestive for a certain event type (e.g., the lack of any location mention correlates with low probability of `Disasters & Accidents` type of an event). Furthermore, different combinations of entity types can be indicative of different event classes. We then detect and generalize entities by their types. In order to select and type the entities mentioned in descriptions we apply Yodie [3] - a named entity recognition and disambiguation (NERD) tool. Finally, we count how many entity types an event description contains (F_5).

Head entity/verb based features. Often the head entity is the main actor in the event, hence, we use its type as an additional feature (F_6). We also extract the head verb which is likely to denote the main action of the event and we map it to VerbNet for obtaining its class (F_7) as the semantic representation of the action of the head entity.

Concept based features. We use Wikipedia as a knowledge base to find similar events to the target event as well as to capture semantic concepts underlying an event description. For mapping descriptions into Wikipedia we employ ESA [7] which outputs Wikipedia articles ranked based on their correspondence to the input text. Many of the returned articles are actually about past events similar to the target event.

Next, we collect categories of the obtained Wikipedia articles. In particular, we fetch all the categories of the top-10 articles given by ESA. Then we use terms from the category names (F_8) and ones from the titles of the top articles (F_9) as additional representation of the target event based on TF-IDF weighting.

[4] https://verbs.colorado.edu/~mpalmer/projects/verbnet.html.

Finally, we combine all the features and perform feature selection to avoid sparsity. In particular, we select k-important ($k = 2,000$[5]) features by using the forests of trees[6].

4 Experimental Evaluation

4.1 Setup

For constructing the feature set for our dataset we have collected 17,503 entities (on average, 2.5 per event), 72,540 Wikipedia articles/concepts (on average, 10 per event) and 116,809 Wikipedia categories (22.5 per event, on average). We trained and tested three kinds of classifiers, SVM with RBF kernel, Naive Bayes Classifier (NB) and Random Forests (RFs), on the nine feature groups in a One-vs-All classifier mode using 10-fold cross-validation. The dimension sizes of LSA and Doc2Vec were set to 300 after experimenting with different numbers on a small held-out dataset.

We compare our approach with classifiers proposed in [4,9]. The former one achieves event classification by training SVM on TF-IDF weighted BOW vectors. The latter is one of the most widely used algorithms for short text classification. That method adds hidden topics into feature vectors in addition to term based features, and trains a MaxEnt classifier.

In addition, for a deeper investigation of our approach we train SVMs separately on each feature group that we collected to analyze how much the particular features and the feature selection can improve the results.

4.2 Discussions of Results

Table 2 compares F-measures of our approaches with that of baselines. SVM equipped with all the features achieves the best results for almost all the classes as well as on the whole dataset. Looking at the micro-average ROC curves (see Fig. 1) for the three classifiers trained on all the feature groups and for the two baselines we see that SVM indeed performs the best among all the compared classifiers. We then focus on SVM in the rest of our analysis.

Looking at Table 2 again we can obtain the detailed analysis of F-measure on the feature group level. The conclusion is that combining all the features improves F-measure for almost all the classes. Especially, the F-measures for `AC`, `BE`, and `ST` are improved over 10% compared with the best results of individual feature groups.

Next, Fig. 2 shows the precision, recall and F-measure per each class. The results are reasonably high (a bit less than 80%) for total Precision, Recall, and F-measures. Weaker results for `HE` and `AC` classes are likely due to relatively small size of training data for these classes as indicated in Table 1.

[5] This value was empirically chosen based on analyzing the results on the small held-out development dataset.

[6] http://scikit-learn.org/stable/index.html.

Table 2. F-measures for SVM obtained when using individual feature groups vs. all features used together for SVM, NB and RFs settings for each class.

Class	SVM with individual feature groups										Proposed methods		
	F_1 ([4])	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	[9]	All+NB	All+RFs	All+SVM
AA	68.2%	10.1%	83.8%	52.5%	28.5%	23.7%	0.0%	69.6%	28.6%	52.3%	50.0%	46.4%	**85.3%**
AC	21.6%	9.1%	41.4%	13.0%	6.2%	0.0%	0.0%	44.8%	7.3%	**79.9%**	70.0%	68.4%	59.7%
BE	36.5%	3.7%	66.2%	21.8%	7.6%	0.0%	0.0%	59.8%	1.4%	73.3%	61.9%	58.2%	**75.5%**
DA	65.5%	22.0%	83.8%	37.5%	1.8%	30.3%	0.0%	68.3%	9.1%	84.4%	64.7%	60.1%	**88.4%**
HE	42.9%	4.4%	54.5%	8.2%	2.2%	3.8%	0.0%	25.9%	3.3%	**89.3%**	72.2%	66.5%	54.0%
LC	42.0%	11.4%	68.3%	33.2%	14.5%	0.0%	0.0%	46.8%	10.2%	65.2%	49.0%	44.4%	**72.0%**
PE	52.7%	15.0%	72.6%	39.0%	20.3%	0.0%	0.0%	61.9%	9.7%	65.5%	58.6%	50.0%	**77.6%**
ST	31.3%	6.8%	58.6%	8.9%	5.0%	0.0%	0.0%	63.6%	0.0%	8.3%	43.0%	44.9%	**71.8%**
S	66.7%	8.2%	85.0%	36.5%	14.4%	0.0%	13.7%	81.8%	10.5%	57.3%	50.3%	52.0%	**89.3%**
Total	*54.5%*	*27.2%*	*74.7%*	*40.9%*	*28.5%*	*11.6%*	*1.0%*	*62.6%*	*46.0%*	*64.0%*	*58.3%*	*54.6%*	***79.7%***

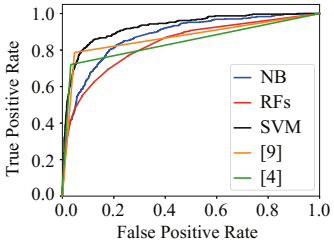

Fig. 1. ROC curves.

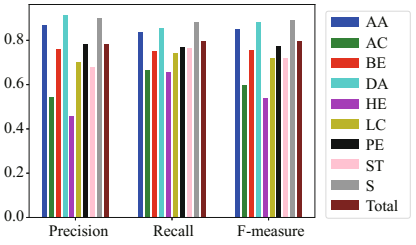

Fig. 2. Results of the proposed approach with SVM.

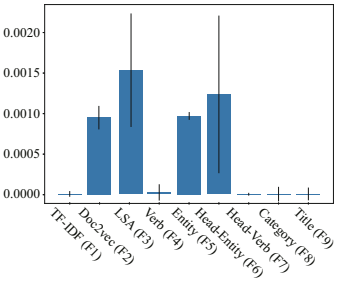

Fig. 3. Feature importance. (Color figure online)

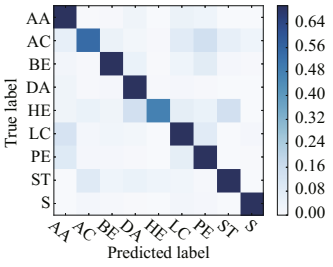

Fig. 4. Confusion matrix on 3,260 randomly sampled events.

In Fig. 3 we show average importance values (blue bars) and standard deviations (black lines) of our features. We can see that LSA and entities (especially the head ones) play importance roles in short event description classification. TF-IDF, verbs, texts of Wikipedia articles' titles and categories were not very important for this task.

Finally, in Fig. 4, we analyze how the classifier mistakes the classes. `AC` tends to be often confused with `PE`, while `LC` and `PE` tend to be often mistaken with

AA. The reason could be that actors in these events are often nations, countries, regions or persons referred to by their nationalities (e.g., "A Japanese scientist"). HE events are sometimes mistakenly classified as ST, likely due to the discoveries in medicine and biology or similar areas. In addition, HE and DA events sometimes occur due to the same trigger. For example, the outbreak of Zika virus in 2016 caused death of many people (HE event) but also the decrease in the population of bees (DA event).

5 Conclusions

It is quite common to briefly refer to past events. Understanding categories of referred events can have many applications including support for building historical analogy models, across-time connection of events/entities or structuring longer text collections such as Wikipedia (e.g., year related articles). In this paper we introduce classification technique for short, retrospective descriptions of events and report satisfactory results over the dataset of 32k event descriptions.

Acknowledgments. This work was supported in part by MEXT Grant-in-Aids (#17H 01828 and #17K12792) and MIC SCOPE (#171507010).

References

1. Au Yeung, C.M., Jatowt, A.: Studying how the past is remembered: towards computational history through large scale text mining. In: CIKM 2011, pp. 1231–1240 (2011)
2. Deerwester, S., Dumais, S.T., Furnas, G.W., Thomas, K.L., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inform. Sci. **41**(6), 391–407 (1990)
3. Gorrell, G., Petrak, J., Bontcheva, K.: Using @Twitter conventions to improve #LOD-based named entity disambiguation. In: Gandon, F., Sabou, M., Sack, H., d'Amato, C., Cudré-Mauroux, P., Zimmermann, A. (eds.) ESWC 2015. LNCS, vol. 9088, pp. 171–186. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18818-8_11
4. Košmerlj, A., Belyaeva, E., Leban, G., Grobelnik, M., Fortuna, B.: Towards a complete event type taxonomy. In: WWW 2015 Companion, pp. 899–902. ACM, New York (2015)
5. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: ICML 2014, vol. 32, pp. 1188–1196, Bejing, China, 22–24 June 2014
6. Lee, U., Liu, Z., Cho, J.: Automatic identification of user goals in web search. In: WWW 2005, pp. 391–400. ACM, New York (2005)
7. Chang, M.W., Ratinov, L.A., Roth, D., Srikumar, V.: Importance of semantic representation: dataless classification. In: AAAI, p. 7 (2008)
8. Nie, L., Wang, M., Zha, Z., Li, G., Chua, T.S.: Multimedia answering: enriching text qa with media information. In: SIGIR 2011, pp. 695–704. ACM, New York (2011)
9. Phan, X.H., Nguyen, L.M., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: WWW 2008, pp. 91–100. ACM, New York (2008)

10. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.: Short text classification in twitter to improve information filtering. In: SIGIR 2010, pp. 841–842. ACM, New York (2010)
11. Sun, X., Wang, H., Yu, Y.: Towards effective short text deep classification. In: SIGIR 2011, pp. 1143–1144. ACM, New York (2011)
12. Zelikovitz, S., Marquez, F.: Transductive learning for short-text classification problems using latent semantic indexing. Int. J. Pattern Recognit Artif Intell. **19**(2), 146–163 (2005)