Multi-label Classification for Past Events

Yasunobu Sumikawa
University Education Center
Tokyo Metropolitan University
Tokyo, Japan
ysumikawa@acm.org

Ryohei Ikejiri
Interfaculty Initiative in Information Studies
The University of Tokyo
Tokyo, Japan
ikejiri@iii.u-tokyo.ac.jp

***Abstract*—Study and analysis of past events can provide numerous benefits. While event categorization has been previously studied, it was usually assigned only one event category to an event. In this work we focus on multi-label classification for past events that is a more general and challenging problem than the previous studies. We categorize them into 13 event categories using a range of diverse features and report micro-average F_1 score is improved approximately by 10% compared with the state-of-the-art algorithm.**

***Index Terms*—Multi-label classification, document classification, history, event**

I. Introduction

Study and analysis of past events can provide numerous benefits, including an enhanced perception of the legacies of the past in the present and enabling learners to make valuable connections through time [4], [17]. One of the goals of imparting recent history education at high schools is to enable students to study how people or organizations in history tried to solve problems in the events. Students can then apply this knowledge to consider creative solutions to social problems in present events [9]. In addition, there are many applications if we correctly understand event descriptions. For example, by being able to tell the categories of mentioned events one could better understand thanks to studying which past event types are mentioned in news articles. 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).

We focus in this work on the problem of *multi-label classification (MLC) for events* that assigns more than one category to each event. For example, if we read the Wikipedia article[1] to know what the 2014 West Africa Ebola outbreak caused in our life, we can see that it killed many both human and nonhuman (environment event), we developed a vaccine (technology event), some researchers report the details and their statistics (academic event), and so on. Tab. I shows other examples of multi-labeled events.

TABLE I
Example events. Our classifier takes descriptions of events; however, We put only short descriptions or names of events to save space in this table. The abbreviated names of categories are used: Reign (Rg), Diplomacy (Dp), War (Wr), production (Pr), Commerce (Cr), Study (St), Religion (Rl), Literature and Thought (LT), Technology (Tc), Popular Movement (PM), Community (Cn), Disparity (Ds) and Environment (En).

Event	Categories
Agnes Chan named UNICEF Regional Ambassador for East Asia and Pacific Region.	Dp, Cn and LT
The World Strikes a Deal on Climate Change.	En
Paris attacks.	Dp, Rg and PM
ISIS Terrorists Strike on Three Continents.	Dp, Rl, Wr and PM
Same-Sex Marriage Debate.	LT and Cn
Ebola outbreak.	En, AC and Tc
The Scottish independence referendum.	Rg, PM and Cn

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. We make an assumption that the context of such descriptions (e.g., surrounding sentences in original text) is not available to cover also the case of standalone descriptions like the lists of significant events in each month of the Wikipedia's Current Portal[3]. Hence we rely only on the event description itself.

To provide sufficient data we use a range of features based on lexical analysis as well as ones based on distributional word representation using neural networks. We use news articles that have one or more than two event categories, and then train our classifiers from the features. There are several labeled event datasets; however, many of them assign only one category to each event. To perform MLC for events, we create a new database and open it on our server (Sec. III).

The contributions of this paper can be summarized as follows:

1) We propose a novel feature selection method.
2) We create a new dataset available online.
3) We conducted evaluations and then confirmed that our classifier achieves approximately 60% in micro-average F_1 score. This result is more 10% better than baselines.

This work was supported in part by MEXT Grant-in-Aids (#17K12792 and #26750076).

[1] https://en.wikipedia.org/wiki/West_African_Ebola_virus_epidemic

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

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

The remainder of this paper is organized as follows: Section II provides summaries of several related works. In Section III, we describe our dataset. We propose our method of feature vector selection in Section IV. Section V describes experimental results. Section VI contains our conclusions.

II. Related Work

Kosmerlj *et al.* propose event categories that are originally defined by Wikipedia editors, and then investigated automatic classification using TF-IDF created from news articles [6]. Several events can be mentioned with a few sentences, such as news articles containing references to related events, historical accounts or biographies. In categorizing short descriptions task, the scarcity of data becomes a more severe problem than long descriptions. To overcome the problem, some studies use context information. Sriram *et al.*'s [16] approach classifies tweets by using author information, url and hashtags of tweets. Nie *et al.* [13] use Naive Bayes classifier equipped with texts, image and video contents for Q&A classification. Lee *et al.* [10] classify queries using user-click behavior to identify user goals in web search. On the other hand, using external information such as Wikipedia resource is also a popular approach. Zelikovitz and Marquez [22] train a classifier with LSA [3] based on Wikipedia data, and Phan *et al.* [15] propose 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) is applied in [19] to map short texts to Wikipedia articles. Sumikawa and Jatowt propose a feature selection method to classify short descriptions of past events [18]. These studies propose classifying event description frameworks; however, they are designed as multi-class classification that assigns only one category to an event.

III. Data Collection

A. Event categories

We use 13 categories: `Reign` (`Rg`), `Diplomacy` (`Dp`), `War` (`Wr`), `production` (`Pr`), `Commerce` (`Cr`), `Study` (`St`), `Religion` (`Rl`), `Literature and Thought` (`LT`), `Technology` (`Tc`), `Popular Movement` (`PM`), `Community` (`Cn`), `Disparity` (`Ds`) and `Environment` (`En`). They are described in [5] as a proposal of an event category list to define the curriculum of teaching history with connecting past and present. These categories are based on definitions of Encyclopedia of Historiography [14]. We show example events for the 13 categories in Tab. I[4].

B. Datasets

In this paper, we use news articles describing events. These articles typically have enough words for classification; however, most of all news articles are assigned categories defined by their companies or organizations. Thus, they are usually different from the above 13 event categories. To train our classifiers, we manually assigned more than one event categories from the 13 ones to several news articles. The assignment processes were done by two Japanese researchers working on history education research and HistoInformatics. They all have Ph. D. degrees; therefore, the dataset is created by experts. We open this new ground truth dataset on our web server[5].

C. Statistics of Dataset

We use news articles included in the Mainichi newspaper articles published in 2012[6]. This dataset includes approximately 100,000 articles for each year. As for 2012, this dataset includes 110,587 articles. We manually select them reporting events, and then we prepared 130 labeled and 9,337 unlabeled news articles as our dataset.

TABLE II
Number of articles per category

Category	Num. of articles	Category	Num. of articles
Cr	45	St	14
Dp	58	Cn	29
Pr	25	LT	15
Rg	43	PM	27
En	14	Tc	20
Rl	26	Wr	21
Ds	18		

We show the number of articles per category in Tab. II. For each event category, there are at least 10 labeled articles. We summarize the statistics of our dataset in Tab. III.

TABLE III
Statistics of dataset

Num. of categories	13
Num. of labeled description	130
Num. of unlabeled description	9,337
Ave. length	887.9
Ave. num. of categories per description	2.7
Ave. num. of description per category	27.3

As our experiment uses Japanese texts, we performed morphological analysis [7] to divide words as no spaces between words in Japanese. We also remove stop words and facilitate stemming are also applied at this time.

IV. Feature Selection

In this section, we describe how our approach creates feature vectors to train classifiers.

A. Word-based features

First, we create TF-IDF vectors (v_1) from all the event descriptions to measure similarity based on their terms.

[4]We use Japanese news articles to evaluate classifications in this paper as described in Sec. V. Even though we did not use the listed example events in the evaluation, we show them to ease understanding what kinds of events can be assigned to from the 13 categories.

[5]http://www.historymining.org/files/13category_events.txt

[6]CD-Mainichi Newspapers 2012 data, Nichigai Associates, Inc., 2012 (Japanese)

B. Semantic-based features

Second, we use all Doc2Vec [8] (v_2), LSA (v_3), and LDA [2] (v_4) to capture latent semantic structures of texts.

C. Noun-based features

Nouns play a key role to distinguish event categories. For example, diplomacy events tend to include names of politicians whereas commerce events frequently mention production items. In the same categories, the nouns tend to have similar semantics. For example, if there are two events *the prime minister Abe Shinzou proposed new trading policies* and *the prime minister Theresa May negotiated trading rules with Japan*, both of the nouns, Abe Shinzou and Theresa May, are politicians leading their countries.

To capture the semantic similarity between nouns, we perform word embedding by Skip-gram model [11]. As this technique assigns vectors to each word where the closer the meaning of them, the greater similarity they indicate, we replace all nouns in event descriptions with their top-k closed words on the vectors. For example, if 5 words *prime*, *minister*, *proposed*, *policies*, and *trading* are the top-5 closest words to *Abe Shinzou* on the vector space, we replace *Abe Shinzou* with the 5 words on event descriptions. We then create TF-IDF vectors (v_5) from the replaced words.

D. Combining Feature Vectors

Finally, we combine all the features, and then perform feature selection to avoid sparsity. Let s_i is a size of the ith feature vector. For each event description, we create 5 feature vectors (v_1, v_2, ..., v_5), and then simply combine them as a feature vector; therefore, the size of a combined feature vector is $s_1 + s_2 + ... + s_5$. For the combined feature vectors, we apply a method of dimensional reduction.

V. Experimental Results

A. Experimental Design

Evaluation criteria. There are several ways to measure performances of MLC in several different points of views. Usually, these performances are measured by two kinds of methods: label-based measures and example-based loss functions [20]. The label-based measures decompose the evaluation with respect to each label whereas the example-based loss functions compute the average differences of the actual and the predicted sets of labels over all examples.

As for the label-based measurement, we use micro- and macro-average precision, recall and F_1 score. These micro-average measurements calculate metrics globally by counting the total true positives, false negatives and false positives. In contrast, the macro-average measurements treat all classes equally; in other words, they compute the metrics independently for each class and then take the average.

In addition, as for multi-label accuracy, we use Jaccard index based measurement. This measurement calculates a score by the dissimilarity between two sets by dividing the difference of the sizes of the union and the intersection of the two sets with the size of the union.

As for the example-based loss functions, hamming loss (HL), ranking loss (RL) and log loss (LL) are popular measurements in MLC. HL calculates the fraction of the wrong labels to the total number of labels. RL means a proportion of pairs of labels which are not correctly ordered. Finally, LL calculates scores from probabilistic confidence. This metric can be seen as cross-entropy between the distribution of the true labels and the predictions. In these measurements, the smaller these scores, the better the performances of the model. We calculate all the above scores by averaging of 5-fold cross-validation.

Parameters. We set the both of the numbers of dimensions of LDA and LSA are 20. For Word2Vec and Doc2Vec, we set 100 as the dimensional size. For creating v_5, we set 5 as k.

Algorithms. We trained our event classifier as SSL (semi-supervised learning) because we must assign more than one event categories to each news article as described in Sec. III; therefore, we perform SSL style event classifier training to reduce the preparation cost. We have implemented two kinds of SSL classifiers: EM algorithm-based classifiers and graph-based ones. These classifiers are listed as follows:

1) Naive Bayes + EM algorithm (NB): We trained Naive Bayes classifier with EM algorithm.
2) Random Forests + EM algorithm (RFs): We trained Random Forests classifier with EM algorithm.
3) SVM (RBF kernel) + EM algorithm (SVM-RBF): We trained SVM whose kernel is an RBF one with EM algorithm.
4) SVM (Linear kernel) + EM algorithm (SVM-Lin.): We trained SVM whose kernel is a linear one with EM algorithm.
5) Label Propagation (LP): LP is a graph-based SSL classification algorithm [23]. This algorithm takes cluster assumption meaning that similar nodes tend to have common labels to calculate scores for assigning categories. This calculation is performed by iteratively multiplying label scores with similarities between nodes.
6) Dynamic LP (DLP): DLP is an extension of LP to take label correlation [21].
7) LP using amendable clamping (LPAC): LPAC is the state-of-the-art algorithm of LP-based algorithm [12]. LPAC is originally designed for label completion task of MLC by emphasizing the cluster assumption; however, this algorithm achieves better than traditional classifiers on a simple MLC task. We use LPAC as a baseline in this study.

We trained the first four classifiers as one-vs-rest classification.

We checked three methods of feature selection as follows:

1) L1 Norm Regularization (L1): This method trains linear model penalized with the L1 norm, and then selects the non-zero coefficients.
2) Random Forests (RFs): This method calculates importance for each feature, and discard irrelevant features according to the values of importance.

3) PCA: This method decomposes a multivariate dataset in a set of successive orthogonal components that explain a maximum amount of the variance.

B. Discussions of Accuracies

First of all, we evaluate how the SSL style training improves micro-average F_1 scores for all classifiers that are trained on all feature vectors combined by L1 based feature selection. We show the results in Figs. 1 and 2. In the both of the two figures, the y axis represents the F_1 score. In Fig. 1 the x axis represents the number of iterations of EM algorithm whereas the x axis of Fig. 2 represents the iteration numbers to train the graph-based algorithms. We can see that NB achieved the best score, approximately 60%, at the second iteration of EM algorithm.

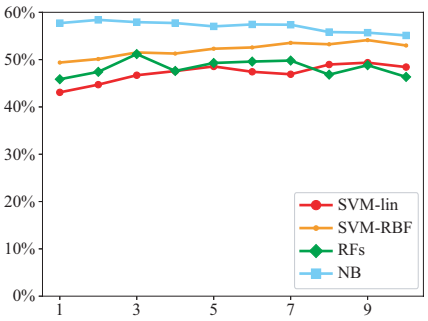

Fig. 1. Micro-average F_1 scores of EM algorithm-based SSL style classification.

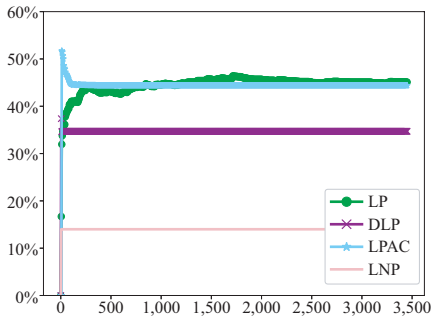

Fig. 2. Micro-average F_1 scores of graph-based classification.

Fig. 3 shows F_1 scores of NB for three different feature selection methods. We can see that L1 based feature selection is the best method; therefore, we show results of classifiers using L1 based feature selection in the following this section.

Next, we show micro-average F_1 scores for all baselines and our approaches in Tab. IV. NB equipped with all the features achieved the best results for almost all the categories as well as on the whole dataset. Thus, we can conclude that combining all the features improves F_1 score for almost all the categories. Especially, the F_1 scores for 9 categories, `Cr`, `Dp`, `Pr`, `Rl`, `St`, `LT`, `PM`, `Tc` and `Wr`, were improved over 10% compared with the best results of individual feature groups. Weaker a result for

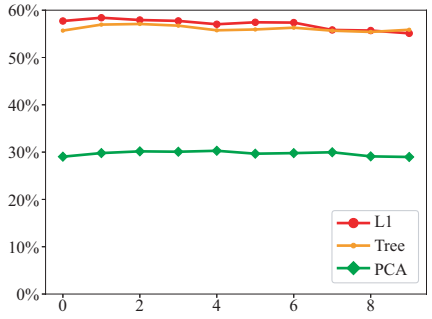

Fig. 3. Micro-average F_1 scores of NB for different algorithms of dimension reduction.

`En` class was likely due to relatively small size of training data for the class as indicated in Tab. II. To better understanding reasons about why 3 categories `Cn`, `Wr` and `Ds` were weak results, we plot the number of co-occurred category pairs in Fig 4. Looking at 2 categories `Cn` and `Wr`, `Dp` is often used with these two categories. Comparing `Dp`, the numbers of articles of the 2 categories are almost half size; therefore, the reason of weak results for the two categories can be considered as the small size of training data. We can also see similar situation for `Ds` descriptions.

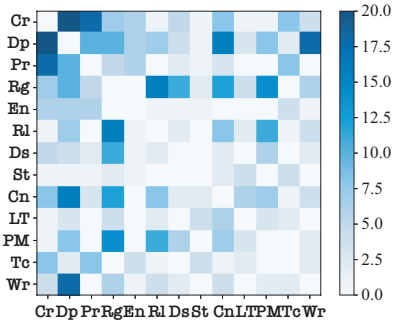

Fig. 4. Co-occurrences of labels.

We next evaluate two other kinds of accuracies (macro-average F_1 (MF) and multi-label accuracy (MA) and three loss scores (HL, LL and RL) in Tab. V. In almost of all measurements, combining all feature vectors improved the scores. Especially, our SVM and LPAC achieved the best scores. In RL measurement, the baseline Doc2Vec achieved the best score. This result indicates that if it is important to decrease RL scores in some application, Doc2Vec based feature vectors can be useful. However, we can conclude that combining all the features improves scores for almost all the categories from the almost of all results of Tabs. IV and V.

Because we observed that our 3 classifiers, NB, SVM-RBF and LPAC, achieved better scores, we check their micro-average precisions and recalls as well as F_1 scores in Figs. 5, 6 and 7. We can see that almost of all results for NB achieved more than 50% for the 3 measurements. Although SVM-RBF and LPAC achieved the best scores in almost of all loss

TABLE IV

F_1 SCORES FOR NB OBTAINED WHEN USING INDIVIDUAL FEATURE GROUPS VS. ALL FEATURES USED TOGETHER FOR NB, RFS AND SVM SETTINGS FOR EACH CLASS. THE BOLD-FACED NUMBERS INDICATE THE BEST ON A PARTICULAR TERMS GIVEN THE METRIC.

	NB with individual feature groups								Proposed methods						
Category	TF-IDF	Doc2Vec	LSA	LDA	Noun	LP	DLP	LPAC	All+NB	All+RFs	All+SVM-RBF	All+SVM-Lin.	All+LP	ALL+DLP	ALL+LPAC
Cr	51.6%	56.3%	55.5%	57.1%	54.9%	36.0%	59.5%	50.4%	70.4%	56.4%	**71.5%**	64.0%	61.0%	60.1%	53.8%
Dp	62.1%	63.7%	63.0%	64.8%	61.9%	66.0%	59.5%	**75.2%**	71.8%	68.8%	68.1%	68.3%	68.4%	59.0%	69.6%
Pr	38.3%	40.1%	36.6%	37.5%	35.0%	20.0%	0.0%	45.0%	**58.5%**	49.8%	47.9%	31.1%	43.9%	43.0%	58.3%
Rg	54.9%	50.2%	54.8%	53.5%	50.6%	52.0%	47.1%	50.1%	54.7%	**59.3%**	57.9%	52.3%	39.0%	47.6%	53.6%
En	16.2%	21.0%	27.1%	24.4%	24.9%	0.0%	0.0%	12.4%	32.0%	**41.3%**	0.0%	0.0%	0.0%	4.1%	0.0%
Rl	40.1%	46.0%	43.4%	34.2%	43.1%	69.5%	20.7%	61.9%	**76.5%**	61.8%	63.3%	56.3%	52.2%	21.0%	71.8%
Ds	31.7%	27.0%	32.2%	26.1%	25.7%	28.8%	10.7%	22.6%	**37.3%**	21.3%	30.0%	8.0%	16.0%	27.0%	23.4%
St	26.0%	32.8%	25.8%	26.5%	27.8%	40.0%	0.0%	50.0%	**69.0%**	63.3%	64.3%	52.9%	6.7%	0.0%	27.4%
Cn	36.2%	34.5%	38.1%	35.6%	**38.3%**	13.6%	32.3%	34.9%	21.9%	23.6%	0.0%	15.7%	22.8%	31.3%	30.1%
LT	24.8%	36.9%	20.0%	38.4%	31.6%	0.0%	0.0%	38.1%	**63.9%**	41.0%	42.7%	13.3%	0.0%	0.0%	0.0%
PM	40.9%	36.5%	37.0%	32.3%	32.7%	35.3%	47.1%	44.7%	**52.4%**	36.9%	0.0%	0.0%	56.2%	47.6%	47.0%
Tc	36.4%	34.4%	34.5%	26.1%	27.7%	19.4%	0.0%	28.0%	56.3%	24.4%	53.4%	**63.3%**	25.2%	21.0%	39.9%
Wr	33.7%	28.1%	33.8%	28.2%	30.5%	39.0%	26.7%	**42.8%**	41.3%	11.7%	0.0%	5.0%	23.1%	25.7%	35.3%
Total	*42.4%*	*41.9%*	*42.2%*	*40.4%*	*40.3%*	*43.3%*	*39.2%*	*49.6%*	***58.4%***	*51.2%*	*54.0%*	*49.4%*	*46.4%*	*38.6%*	*51.5%*

TABLE V

SCORES OF MACRO-AVERAGE F_1 (MF), MULTI-LABEL ACCURACY, (MA), HAMMING LOSS (HL), LOG LOSS (LL), RANKING LOSS (RL).

	MF	MA	HL	LL	RL
TF-IDF	44.8%	79.6%	0.2035	6.8795	0.4442
Doc2Vec	72.5%	79.4%	0.2059	0.9671	**0.1400**
LSA	67.3%	77.5%	0.2248	0.5622	0.1993
LDA	65.2%	79.4%	0.2059	0.4721	0.2286
Noun	55.1%	81.1%	0.1893	6.1565	0.3359
LP	32.3%	26.4%	0.2224	6.4113	0.2796
DLP	23.3%	25.3%	0.4740	6.7732	0.347
LPAC	62.0%	74.2%	0.2580	0.7622	0.2140
All+NB	71.9%	80.0%	0.2	1.3951	0.1503
All+RFs	69.3%	80.5%	0.1952	0.8406	0.2871
All+SVM-RBF	71.1%	**82.2%**	**0.1781**	**0.4261**	0.1887
All+SVM-Lin.	68.0%	81.1%	0.1887	0.4579	0.2122
All+LP	62.4%	74.9%	0.2515	1.5260	0.3129
All+DLP	29.8%	23.8%	0.6041	6.7671	0.3276
ALL+LPAC	**75.1%**	65.1%	0.2488	0.9335	0.2487

functions, some scores of micro-average precision, recall and F_1 were quite low, especially, scores for `En` (`Environment`) in both of SVM-RBF and LPAC, `Cn` (`community`), `PM` (`Popular Movement`) and `Wr` (`War`) in SVM-RBF and `LT` (`Literature and Thought`) in LPAC were 0%. These results indicate that NB is the best algorithm in average.

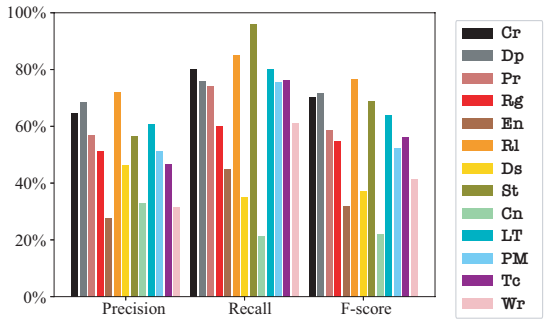

Fig. 5. Micro-average precision, recall andF_1 scores for All+NB.

In Fig. 8 we show average importance values (blue bars) and standard deviations (black lines) of our features. We can see that the noun-based feature was the most important in multi-

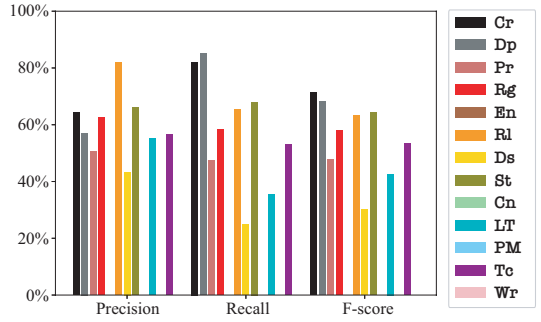

Fig. 6. Micro-average precision, recall and F_1 scores for All+SVM-RBF.

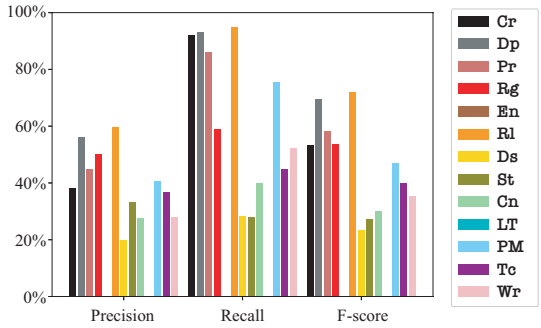

Fig. 7. Micro-average precision, recall and F_1 scores for All+LPAC.

label event description classification. TF-IDF was a little bit important feature. In contrast, all of the semantic base features were not very important for this task. We believe that once we increase the number of labeled descriptions, importances for the semantic base features will be increased.

Finally, we try to analyze what and why our classifier performed mis-predictions. At the beginning, we show what categories were wrongly assigned to events by our classifier (All+NB) in Fig. 9. We can see that `Wr` was often assigned to `Cn` events wrongly. This is because both of the two categories tend to mention locations. For example, as the community-related events, some countries or regions are mentioned with

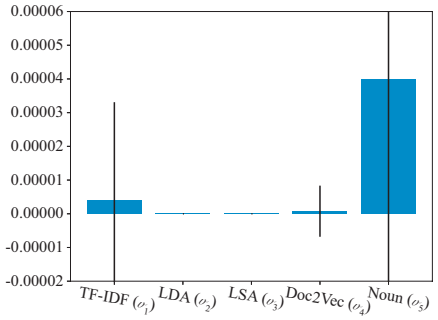

Fig. 8. Feature importances.

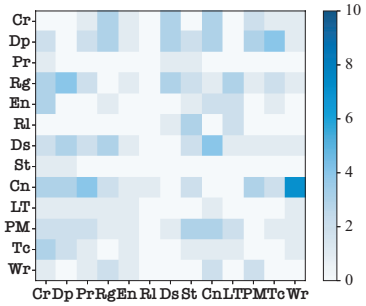

Fig. 9. Wrongly assigned categories. The x axis represents the number of categories that are wrongly assigned to if a category of y axis is not assigned to.

issues of economic or political policies for the places. On the other hand, as for the war event, several countries are mentioned as the main actors of the events.

Next, we count the number of categories that are attached in the test data but our classifier did not assign in Fig. 10. We can see that several test data that are attached both of two categories `Cr` and `Dp` tend to be assigned one of them.

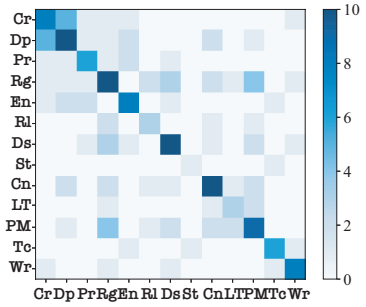

Fig. 10. Missed categories. The x axis represents the number of categories that are correct but are not assigned to if a category of y axis is not assigned to.

VI. Conclusions

Understanding categories of 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 a classification technique for multi-labeled descriptions of events. We showed that our technique could improve micro-average F_1 scores by approximately 10%. For this evaluation, we created a new ground truth dataset, and open it on our web server.

Future work will identify *how the accuracies can be improved by increasing labeled descriptions*. Our current dataset is relatively small; it includes only 130 labeled and 9,337 unlabeled descriptions. We will add more labeled data from other datasets such as The New York Times.

References

[1] Au Yeung, C.m., Jatowt, A.: Studying how the past is remembered: Towards computational history through large scale text mining. CIKM '11, pp. 1231–1240. ACM, New York, NY, USA (2011)
[2] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
[3] Deerwester, S., T. Dumais, S., W. Furnas, G., Thomas K., L., Harshman, R.: Indexing by latent semantic analysis. J. Amer. Soc. Inform. Sci. **41**(6), 391–407 (1990)
[4] Harris, R., Rea, A.: Making history meaningful: helping pupils to see why history matters. Teaching History **125**, 28–36 (2006)
[5] Ikejiri, R., Sumikawa, Y.: Developing world history lessons to foster authentic social participation by searching for historical causation in relation to current issues dominating the news. Journal of Educational Research on Social Studies **84**, 37–48 (2016). (in Japanese)
[6] Kosmerlj, A., Belyaeva, E., Leban, G., Grobelnik, M., Fortuna, B.: Towards a complete event type taxonomy. pp. 899–902. WWW '15 Companion, ACM, New York, NY, USA (2015)
[7] Kudo, T., Yamamoto, K., Matsumoto, Y.: Applying conditional random fields to japanese morphological analysis. EMNLP '04, pp. 230–237
[8] Le, Q., Mikolov, T.: Distributed representations of sentences and documents. pp. 1188–1196. ICML'14, Bejing, China (2014)
[9] Lee, P.: Historical literacy: Theory and research. International Journal of Historical Learning, Teaching and Research **5**(1), 25–40 (2005)
[10] Lee, U., Liu, Z., Cho, J.: Automatic identification of user goals in web search. WWW '05, pp. 391–400. ACM, New York, NY, USA (2005)
[11] Mikolov, T., Yih, W.t., Zweig, G.: Efficient estimation of word representations in vector space. NAACL'13 (2013)
[12] Miyazaki, T., Sumikawa, Y.: Label propagation using amendable clamping. IUI'18 Workshop on WII (2018)
[13] Nie, L., Wang, M., Zha, Z., Li, G., Chua, T.S.: Multimedia answering: Enriching text qa with media information. SIGIR '11, pp. 695–704. ACM, New York, NY, USA (2011)
[14] Ogata, I., Kato, T., Kabayama, K., Kawakita, M., Kishimoto, M., Kuroda, H., Sato, T., Minamizuka, S., Yamamoto, H.: Encyclopedia of historiography. koubundou (1994). URL http://ci.nii.ac.jp/ncid/BN10236869
[15] Phan, X.H., Nguyen, L.M., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. WWW '08, pp. 91–100. ACM, New York, NY, USA (2008)
[16] Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.: Short text classification in twitter to improve information filtering. SIGIR '10, pp. 841–842. ACM, New York, NY, USA (2010)
[17] Staley, D.J.: A history of the future. History and Theory **41**, 72–89 (2002)
[18] Sumikawa, Y., Jatowt, A.: Classifying short descriptions of past events. pp. 729–736. ECIR'18 (2018)
[19] Sun, X., Wang, H., Yu, Y.: Towards effective short text deep classification. SIGIR '11, pp. 1143–1144. ACM, New York, NY, USA (2011)
[20] Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining Multi-label Data, pp. 667–685. Springer US, Boston, MA (2010)
[21] Wang, B., Tu, Z., J.K., T.: Dynamic label propagation for semi-supervised multi-class multi-label classification. ICCV '13, pp. 425–432 (2013)
[22] Zelikovitz, S., Marquez, F.: Transductive learning for short-text classification problems using latent semantic indexing. International Journal of Pattern Recognition and Artificial Intelligence **19**(2), 146–163 (2005)
[23] Zhu, X.: Semi-supervised learning with graphs. Ph.D. thesis, Pittsburgh, PA, USA (2005). AAI3179046