ECIR' 18, 27 March 2018 @Grenoble, France

Classifying Short Descriptions of Past Events

Yasunobu Sumikawa
Tokyo University of Science

Adam Jatowt
Kyoto University

Introduction

Past events can be referred with short texts.
Ex.: chronological ordered lists.

- September 7 – Israel becomes the 33rd member of the OECD.[56]
- September 30 – Germany pays war reparations for World War I.

Contributions:

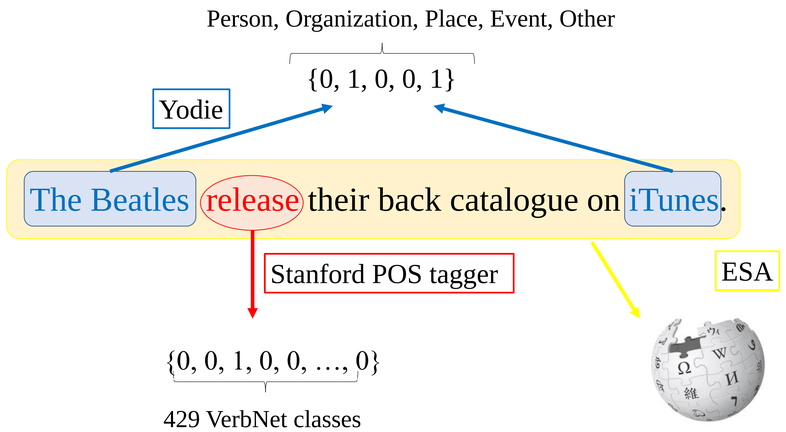
1. Identifying what features are useful for this task.
2. Our classifier is the best for micro-avg. F-score.
3. This is useful for constructing thematic timelines or event lists (e.g., list of disasters/accidents in Asia, timeline of armed conflicts in USA).

Event Classes and Example Descriptions

1. **Armed Conflicts and Attacks** (**AA**, 8886 events),
   Ex. "*Bombs across Iraq detonate, killing 18 people*."
2. **Arts and Culture** (**AC**, 1800 events),
   Ex. "*The Beatles release their back catalogue on iTunes.*"
3. **Business and Economy** (**BE**, 2517 events),
   Ex. "*Brazil's economy falls into recession*."
4. **Disasters and Accidents** (**DA**, 4961 events),
   Ex. "*A bus crashes into a ravine in Tibet, killing at least 44 people*."
5. **Health and Environment** (**HE**, 487 events),
   Ex. "*The number of Zika virus infected in Singapore rises above 40*."
6. **Law and Crime** (**LC**, 4984 events),
   Ex. "*The Constitutional Council of France upholds a ban on fracking*."
7. **Politics and Elections** (**PE**, 5517 events),
   Ex. "*Voters in Costa Rica go to the polls for a general election.*"
8. **Science and Technology** (**ST**, 1066 events),
   Ex. "*Iran successfully puts the Fajr satellite in orbit using a Safir-B1 rocket.*"
9. **Sport** (**S**, 2400 events)
   Ex. "*The Winter Olympics in Sochi, Russia officially concludes.*"

Method

Feature Vector Creation



We combine all the 9 feature vectors: TF-IDF, Doc2Vec, LSA, Verb, Entity, Head-Verb, Head-Entity, Wikipedia Category, Titles of Wikipedia articles.

Dimension Reduction

To reduce sparsity, we select k-important (k = 2,000) features by using the forests of trees.

Statistics of Dataset

Tab.1 Statistics

Num. of entities	17,503
Num. of concepts	2,540
Num. of Wiki. Category	16,809
Num. of typed events	32,618
Timespan range	2010/1/1 ~ 2016/12/17

* Current events portal of Wikipedia.
Average lengths: 25 words.

Evaluations

Baselines

TF-IDF + SVM: SVM on TF-IDF weighted BOW vectors designed for long text.
MaxEnt: Entropy-based classifier that is one of the most commonly used method for short text.

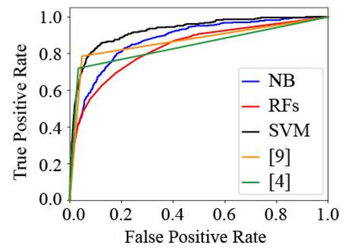
Results

Tab.2 Micro-average F-scores for baselines

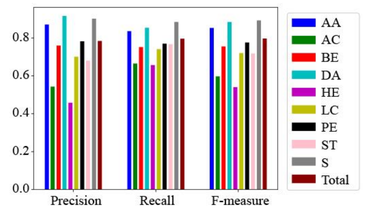
TF-IDF + SVM	MaxEnt
54.5%	64.0%

Tab.3 Micro-average F-scores for our approaches

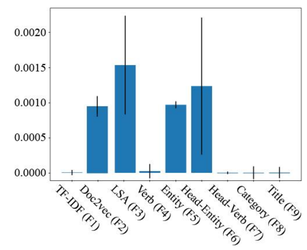
All+Naïve Bayes	All+RFs	All+SVM
58.3%	54.6%	*79.7%*



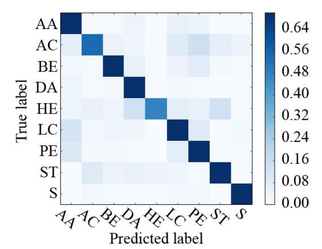
(a) Micro-average ROC curves



(b) Accuracies for All+SVM



(c) Feature importance



(d) Mis-predicted by All+SVM

Conclusions

We introduce classification technique for short, retrospective descriptions of events and report satisfactory results (F-score of about 0.8) over the dataset of 32k event descriptions.