Overview of SYDNEY System for TAC KBP 2015 Event Nugget Detection

Shang Chun Sam Wei
School of Information Technologies
University of Sydney
NSW 2006, Australia
swei4829@uni.sydney.edu.au

Igor Korostil
Capital Markets CRC
55 Harrington Street
NSW 2000, Australia
eeghor@gmail.com

Ben Hachey
School of Information Technologies
University of Sydney
NSW 2006, Australia
ben.hachey@gmail.com

Abstract

We describe the SYDNEY submission to the TAC 2015 event nugget detection shared task. Development experiments explore the contribution of features aimed at improving generalisation. They indicate that Brown cluster, Nomlex and WordNet features are complementary with more improvement from WordNet features. Final submissions differ in the number of negative examples ued for training trigger detection, with 10% subsampling resulting in our best f-score of 51.97.

1 Introduction

Event trigger detection challenges include limited training data, difficult generalisation to unseen data and a highly skewed class distribution. Many previous approaches use a pipelined approach of event trigger detection followed by event type classification (Ahn, 2006; Chen and Ji, 2009) We follow previous work in formulating the problem as a token-level classification task. We use BIO1 labelling to minimise sparsity. And we subsample negative examples to manage class imbalance, preventing the classifier from having a detrimental bias.

To address low recall, we evaluate several features aimed at improving generalisation. These are based on Brown clusters, Nomlex and WordNet. Development experiments suggest that these are complementary. WordNet is the strongest feature by far, improving f-score by 14 points over a classifier with all other features combined. Final submissions include systems trained with 5%, 10% and 15% subsampling of negative examples. 10% performs best with precision of 66.77, recall of 42.53 and f-score of 51.97. The system's overall ranking is fifth.

2 Approach

We built our initial system with a ME classifier. Features include lexical and semantic information from the current token and surrounding context.

The event trigger could be a phrase and requires encoding scheme to represent the chunks. We experimented with Begin Inside Outside 1 (BIO1) and Begin Inside Outside 2 (BIO2) encodings (Sang and Veenstra, 1999) – see table 1. Preliminary results showed little impact on accuracy. However, one of the issues with this task is data sparsity. Some event subtypes have few observations in the corpus. BIO2 encoding increases the total number of categories for the dataset. Thus make the data sparsity issue worse. Therefore we decided to use the BIO1 encoding for the rest of the experiment.

Word	BIO1	BIO2
He	O	O
has	O	O
been	O	O
found	I-Justice.Convict	B-Justice.Convict
guilty	I-Justie.Convict	I-Justice.Convict
for	O	O
the	O	O
murder	I-Life.Die	B-Life.Die
.	O	O

Table 1: BIO1 and BIO2 encoding comparison. "O" represents no event.

The other issue we found is that the data is very unbalanced. Most of the tokens are not event triggers. To overcome this issue, we have tried various sub-sampling of the negative observations (none event tokens). We found that randomly sampling 10% of the negative examples for training works well here.

2.1 Out-of-vocabulary Issue

One of the issue with event trigger detection is data sparsity and out-of-vocabulary (OOV) issue. At token level, there are 1,700 events in the development set. There are 456 (26.82%) tokens not observed in the training set. Further there are 197 (11.59%) tokens only appear in training set twice or less.

For example, below is a *Life.Die* event in the development set that is not in the training set.

*Tamil Self-***Immolation**

This is one of the challenge for event trigger detection. It is evident in our results that the precision is much higher than the recall. The low recall still remain as the bottleneck.

2.2 Features

The features are divided into four different groups. They are:

Feature set 1 (FS1): Basic features including following.

- Current token: Lemma, POS, named entity type, is it a capitalised word.
- Within the window of size two: unigrams/bigrams of lemma, POS, and name entity type.
- Dependency: governor/dependent type, governor/dependent type + lemma, governor/dependent type + POS, and governor/dependent type + named entity type.

Feature set 2 (FS2): Brown cluster trained on Reuters Corpus Volume 1 (RCV1) corpus[1] with prefix of length 11, 13 and 16.

Feature set 3 (FS3): Base form of the current token extracted from Nomlex[2].

Feature set 4 (FS4): 1. WordNet features including hypernyms and synonyms of the current token.

3 Experimental Setup

3.1 Dataset and Evaluation Metric

TAC 2015 dataset (LDC2015E73) is used for the experiment. The corpus has total 158 documents with two genres: 81 newswire documents and 77 discussion forum documents.

The performance is measured using the TAC 2015 event mention scorer.[3] The F1 score is computed using precision P and recall R for the given true positives TP, the number of system mentions N_S, and the number of gold mentions N_G.

$$P = \frac{TP}{N_S}; R = \frac{TP}{N_G}; F1 = \frac{2PR}{P+R}$$

To measure the event trigger boundary and label performance, we use the micro average "mention_type" results from the scorer. An event trigger is counted as correct only if the boundary, the event type and the event subtype are all correctly identified.

The dataset is split into 80% for training (126 documents) and 20% for development (32 documents).

3.2 Event Trigger Detection

The preprocessing steps take raw data and perform tokenisation, sentence split, POS tagging, name entity recognition, constituency parse and dependency parse using Stanford CoreNLP.[4]

We used supervised methods to detect event triggers and classify realis. Scikit-learn (Pedregosa et al., 2011) was used to train our ME classifiers. A ME classifier was trained to detect and label the events. The Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm was used as the solver for the ME model. We also trained a separate ME classifier to label realis.

One disadvantage with this approach is the model will miss double tagged triggers. TAC 2015 shared task is different from previous event tasks such as TAC 2014 and Automatic Content Extraction (ACE) 2005. It is possible for a token to have multiple events. This is common with Conflict.Attack and Life.Die events. There are about five percent such tokens. Our model will only give one label per token. Hence miss the double tagged events.

[1] http://metaoptimize.com/projects/wordreprs/

[2] http://nlp.cs.nyu.edu/nomlex/

[3] Scorer available at http://hunterhector.github.io/EvmEval/

[4] http://nlp.stanford.edu/software/corenlp.shtml

3.3 Feature Set Evaluation

We perform a cumulative analysis to quantify the contribution of different feature sets to the classifier's performance. Table 2 shows the impact of each feature set on the model. The feature set 2 (Brown cluster) helped with recall and precision. The recall is further boosted by feature set 3 (Nomlex) and feature set 4 (WordNet). The most contribution is from feature set 4 that has increased recall by 16.89%.

System	P	R	F1
ME FS1	54.87	20.68	30.04
ME FS1+FS2	57.00	24.68	34.45
ME FS1+FS2+FS3	59.68	26.59	36.79
ME FS1+FS2+FS3+FS4	61.03	43.48	**50.78**

Table 2: Feature comparison on "mention_type" performance.

4 Results

We submitted three runs:

Run 1: Sub-sampling 10% negative observation.

Run 2: Sub-sampling 5% negative observation.

Run 3: Sub-sampling 15% negative observation.

Table 3 shows the "mention_type" results from each run and compare to the top rank system. The metrics are from scorer's "mention_type" micro average. The Run 1 has the highest F-score ranked sixth in this metric category. However, Run 2 has the highest recall and Run 3 has the highest precision. We find sub-sampling negative observations at 10% gives a good balance of precision and recall.

System	P	R	F1
Rank #1	75.23	47.74	58.41
Run 1	66.77	42.53	**51.97**
Run 2	56.89	**47.70**	51.89
Run 3	**70.29**	39.94	50.94

Table 3: System "mention_type" micro average performance comparison.

The evaluation results (51.97 F-score) match closely to our development results (50.78 F-score). All the results have higher precision (56.89% - 70.29%) than recall (42.53% - 47.70%).

Table 4 shows the combined performance of the two classifiers for "mention_type" and "realis_status". Both Run 1 (10%) and Run 2 (5%) has similar F-score. The Run 2 system ranked fifth overall. Again 10% gives reasonable trade off between precision and recall.

System	P	R	F1
Rank #1	56.98	36.16	44.24
Run 1	50.30	32.04	39.14
Run 2	43.12	**36.16**	**39.33**
Run 3	**53.13**	30.19	38.50

Table 4: System overall (mention_type+realis_status) micro average performance comparison.

5 Conclusion

We built an event trigger detection system and evaluate the performance using TAC 2015 corpus. We analysed the effectiveness of the features and the impact on classifier performance. Brown cluster information increased recall and precision slightly. Nomlex further improve the results. WordNet features give significant boost to the performance with 13.99% absolute F-score. Finally, different sub-sampling of negative observations give us trade off between precision and recall.

Acknowledgements

Sam Wei is funded by the Capital Markets Cooperative Research Centre. Ben Hachey is the recipient of an Australian Research Council Discovery Early Career Researcher Award (DE120102900)

References

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning About Time and Events*, ARTE '06, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Comput. Linguist.*, 18(4):467–479, December.

Zheng Chen and Heng Ji. 2009. Language specific issue and feature exploration in chinese event extraction. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, May 31 - June 5, 2009, Boulder, Colorado, USA, Short Papers*, pages 209–212.

Anup Kumar Kolya, Asif Ekbal, and Sivaji Bandyopadhyay. 2011. A hybrid approach for event extraction and event actor identification. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, and Nicolas Nicolov, editors, *RANLP*, pages 592–597. RANLP 2011 Organising Committee.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *ACL (1)*, pages 73–82.

Catherine Macleod, Ralph Grishman, Adam Meyers, Leslie Barrett, and Ruth Reeves. 1998. Nomlex: A lexicon of nominalizations. In *In Proceedings of Euralex98*, pages 187–193.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Erik F. Tjong Kim Sang and Jorn Veenstra. 1999. Representing text chunks. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, pages 173–179, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joseph Turian, Dpartement Dinformatique Et, Recherche Oprationnelle (diro, Universit De Montral, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semisupervised learning. In *In ACL*, pages 384–394.