Eliminating Incorrect Events from Large-Scale Event Networks by Trigger Word Clustering and Pruning

Farrokh Mehryary[1,2], **Suwisa Kaewphan**[1,2,3], **Kai Hakala**[1], **and Filip Ginter**[1]
[1]Dept. of Information Technology, University of Turku, Finland
[2]The University of Turku Graduate School (UTUGS), University of Turku, Finland
[3]Turku Centre for Computer Science (TUCS), Turku, Finland
farmeh@utu.fi sukaew@utu.fi kahaka@utu.fi ginter@cs.utu.fi

Abstract

In this short paper, we investigate hierarchical clustering of event triggers in the EVEX large-scale event resource. As the primary application, we utilize the clustering to identify incorrect trigger event words and subsequently eliminate events extracted with these triggers. We evaluate the method on the BioNLP 2011 and 2013 Shared Task test sets and show that the method can further increase the precision and F-score of the winning system of the 2013 BioNLP Shared Task on Event extraction.

1 Introduction

The 2009–2013 series of BioNLP Shared Tasks on Event Extraction (Nédellec et al., 2013) have given raise to a number of event extraction systems, several of which have been applied at a large scale (Van Landeghem et al., 2013a; Gerner et al., 2012). This typically means the full set of PubMed abstracts and PubMed Central full text articles. Having been trained on the narrow-domain, carefully selected Shared Task training data, the performance of the systems when faced with the topically highly varied PubMed documents — the vast majority of which do not deal with events among genes and proteins — thus requires a further study.

Already a casual inspection of the EVEX resource reveals occasional occurrences of obviously incorrect events in out-of-domain documents. Further, Van Landeghem et al. (2013b) have studied the output of the event extraction systems on general domain data in further detail. In this short paper, we focus specifically on the event triggers, with the objective of automatically identifying those that are obviously incorrect.

2 Data

This study is based on the EVEX resource (Van Landeghem et al., 2013a) containing 40,190,858 events of 24 different types such as *binding, positive-regulation, negative-regulation, and phosphorylation*. These events are extracted using the Turku Event Extraction System (TEES) (Björne et al., 2012) from 6,392,824 PubMed abstracts and 383,808 PubMed Central Open Access (OA) full-text articles that were published up to 2012 and which contain at least one gene/gene-product mention.

The extraction of every event is based on the recognition of an occurrence of a trigger word in the underlying sentence. A single unique trigger word, such as *modify*, may have a number of occurrences in the data, acting as a trigger for many events. It is important to note that these events may be of different types. For instance the trigger word *expression* acts as a trigger for both *gene-expression* and *transcription* events, depending on the context.

In total, there are 137,146 unique event triggers not comprised purely of numbers and not containing unicode special characters. As expected, the vast majority of events in EVEX correspond to a small number of highly frequent trigger words. For example, there are only 3,391 trigger words with frequency above 300 (i.e. corresponding to at least 300 event occurrences), but these words account for full 97.1% of all events in EVEX. Consequently, when the aim is to increase the precision of the events in EVEX by automatic recognition of *incorrect* trigger words and eliminating them, focus should be centered on highly frequent trigger words instead of the rare ones. Accordingly, we decided to concentrate on these 3,391 top most frequent trigger words. Having this limited number of trigger words of focus also made it possible to manually inspect the hierarchical clustering tree discussed in the following sections. Table 1 shows more detailed statistics regarding the distribution of event trigger words.

Trigger Word Frequency (at least)	EVEX Events Coverage %	Number of Trigger Words
100	98.4%	6339
200	97.6%	4263
300	97.1%	3391
400	96.6%	2880
500	96.3%	2538

Table 1: Distribution of triggers and their associated event percentages.

The event type distribution of any particular trigger word can be characterized with a 24 dimensional vector, in which each element corresponds to the proportion of a particular event type, among all events that are extracted based on that particular trigger word. While some trigger words are very pure, where for instance *bind* is 99.92% associated with the *binding* event type, this is often not the case.

3 Method

Among the trigger words, we will target those which are obviously incorrect, regardless of their context. These include, for example, gene/protein/chemical names, author names or names such as *hospital*. One of the objectives of this study is thus to develop a method that can *automatically* categorize the trigger words so as to eliminate the obviously incorrect trigger words, thus increasing the precision of the event extraction systems without impacting their recall.

Another interesting aspect when studying the trigger words is to build a general overview of all of the trigger words according to the 24 different event types and to study whether there exist groups/subgroups of related trigger words which would allow us to define sub-types of the 24 event types. Of specific interest will be studying the groups/sub-groups before and after eliminating incorrect trigger words.

In the following, we propose a 4-step method which is based on hierarchical/agglomerative clustering of the 3,391 top-most frequent trigger words and building, analyzing and pruning the resulting binary tree.

3.1 Hierarchical Clustering of Trigger Words

As the first step, we induced a vector space representation for the trigger words, and used these vectors to hierarchically cluster the triggers. *Cosine similarity* was used as the clustering metric and the *ward variance minimization algorithm* was used for calculating the distances between the newly formed clusters. To build the vector space representations, we used the *word2vec* method of distributional semantics recently introduced by (Mikolov et al., 2013) and previously applied in the biomedical domain by (Pyysalo et al., 2013). The *word2vec* method comprises a simplified neural network model with a linear projection layer and a hierarchical soft-max output prediction layer. The input layer has the width of the vocabulary and the projection layer has the width of the desired dimensionality of the vector space representation. Upon training, the weight matrix between the input and the projection layer constitutes the word vector space embeddings. The network can be trained in several different regimes, but in this work we use the skip-gram architecture, whereby the network is trained to predict nearby context words, given a single focus word at the center of a sliding window context.

We trained the *word2vec* model on the lowercased texts of all PubMed titles and abstracts and all PubMed Central OA full articles that contain at least one gene/gene-products (*ggps*) mentions extracted from the EVEX resource. All *ggp* mentions in the texts were replaced with the *"ggp"* placeholder and all numbers with the *"num"* placeholder to densify the text.

An initial experiment in hierarchical clustering of the top 100 most frequent trigger words revealed that on one hand many coarse/fine grained sub-clusters were formed in a way that each sub-cluster contained trigger words with biologically similar meaning. Many sub-clusters could be clearly associated with a unique event type. However, on the other hand, many trigger words were clustered together incorrectly, especially for the common *positive-regulation* and *negative-regulation* types (e.g. *increase* and *decrease*) because they have a high similarity in the vector space representation.

To address this issue, we add trigger/event type association information as additional dimensions to the word vectors, thereby affecting the clustering to more closely conform to the event types. To obtain reliable event type distribution for the trigger words, we used the BioNLP Shared Task 2011 (ST'11) *training* and *development* sets (Kim et al., 2011). Out of the 1,447 unique trigger words in this data, 995 were single-token trigger words and of these, 828 are actually among the top 3,391 most frequent EVEX trigger words. For these 828 triggers, we append a 24-dimensional normalized event type distribution vector to their *word2vec*-based vectors (the vectors for the remaining 2,563 triggers for which a reliable event type information could not be obtained are simply padded with 24 zeroes). Reclustering with the modified vectors, we noticed that *positive-regulation* and *negative-regulation* trigger words were no longer clustered together, obtaining more meaningful clusters w.r.t. the task at hand.

3.2 Event Type Vectors of Sub-Clusters

In this step, event type vectors for all nodes of the binary cluster tree were calculated. For each leaf of the tree (i.e., a trigger word), its corresponding trigger/event type vector was calculated based on the counts of the occurrence of its respective events in the EVEX, and for each intermediate node of the tree (i.e., a sub-cluster), its respective event type vector was calculated by adding trigger/event type vectors of all triggers that belong to it.

Using this information, it is possible to inspect how the tree is organized and whether and how its different branches represent different event types. For example, by checking which element in a sub-cluster's event type vector has the maximum value, we could tell what is the event type that this sub-cluster is mostly associated with and the level of purity of that cluster. For example, while one sub-cluster can be 98% *binding* and is thus to a large extent *pure*, another cluster can be 43% *gene_expression* and cannot be assigned a single predominant type.

3.3 Identifying *Possibly Incorrect* Trigger Words

In this step, we prepared a list of *safe* or *supposedly to be correct* trigger words and regarded the remaining trigger words in our set of all trigger words of focus as *possibly incorrect*. This was necessary for the next step: pruning the tree and finding the final list of *incorrect* trigger words. As discussed in Section 3.1, by analyzing the ST'11 training and development sets, we obtained a list of 995 unique single-token trigger words. Out of the 3,391 EVEX trigger words in our list, 828 were among these. However, our list contains many other trigger words that could not be directly found in the ST'11 sets, but variations of them or variations of their parts could. For instance, *processing* was in our list, while *processed* was in the ST'11 sets. As another example, the word *co-regulation* was in our list, but *regulation* is present in the ST'11 sets.

We therefore performed the following preprocessing steps on the trigger words in both the EVEX and ST'11 sets:

1. We split each trigger word based on occurrences of the following characters: {'-' , '.' , '_' , '/'}. For each trigger word, each of its split parts was saved if *all* of the following conditions were met: (1) The part length was greater than one, (2) The part was not a number, (3) The part was not in the set of following words: {'co' , 'up' , 'down' , 're' , 'non' , 'over' , 'self'}.
2. We lemmatized all the trigger words, and all of their parts, using the *BioLemmatizer* software (Liu et al., 2012) which is specifically developed for the biomedical domain, and recorded all produced lemmas for each trigger word.

After the preprocessing, all of the 828 trigger words that could be directly found in the ST'11 sets were regarded as *safe*. For the rest, the following matching procedure was performed: A trigger word in our list was regarded as *safe* if its exact form, or one of its parts, or one of the lemmas of its parts could be found in the sets of the ST'11 words, or their parts or part lemmas. Otherwise, the trigger word was regarded as *possibly incorrect*. By performing the abovementioned approach, 602 trigger words were identified and added to the list of *safe* trigger, resulting in a list of 1,430 *safe* trigger words. The 1,961 remaining triggers were regarded as *possibly incorrect*. Table 2 shows some example words from EVEX triggers in our list that were matched against ST'11 trigger words, parts, or lemmas.

EVEX Trigger Word	ST'11-Trigger Word/Part/Lemma
co-transcribed	transcribed
calcium-induced	induced
co-immunoprecipitates	immunoprecipitate
downregulating	downregulate
recognise	recognize
preceding	precede
analyzing	analyse

Table 2: Example of how trigger words from EVEX were matched against ST'11 *exact trigger words* or their *corresponding parts*, or *the lemmas of their parts*.

3.4 Pruning the tree

Pruning the tree was done using the list of *possibly incorrect* trigger words in three steps: 1) Processing all leaves of the tree: If a trigger word exists in the list of *possibly incorrect* trigger words, its corresponding leaf was marked as *unsafe*, otherwise it was marked as *safe*; 2) Processing all intermediate nodes of the tree: If *all* of the children of an intermediate node were marked as *unsafe*, this node was marked as *unsafe* as well, otherwise it was marked as *safe*; 3) Pruning the tree: All of the descendants and leaves of the intermediate nodes that were marked as *unsafe*, were deleted from the tree. Respective trigger words of the deleted leaves were added to the list of *incorrect* trigger words.

There is one important aspect in the pruning algorithm: Because the tree is binary, not all of the trigger words that are in the list of *possibly incorrect* trigger words were added to the list of *incorrect* trigger words, because if such a trigger word was clustered near a *safe* trigger word, it was not considered as an *incorrect* trigger word and remained in the tree. By doing this, deletions were propagated to the upper levels nodes of the tree, only if all of the participating leaves were recognized as *incorrect*.

After pruning, event type vectors for all intermediate nodes of the tree were recalculated so that we could compare the tree before and after pruning.

4 Evaluation

4.1 Event Filtering

We evaluate the impact of trigger pruning on event extraction using the official test sets of the BioNLP'11 and '13 GENIA Event Extraction (GE) shared tasks. As the basis we consider the outputs of the *TEES system* entry (Björne et al., 2012; Björne and Salakoski, 2013) in 2011 (3rd place) and in 2013 (2nd place) GE tasks and, for the 2013 shared task, also the winning *EVEX* entry (Hakala et al., 2013). We prune the outputs of these systems by removing events whose trigger words are identified as incorrect using the abovementioned algorithm and evaluate the resulting pruned set of events using the official evaluation services of the respective Shared Task on the held-out test sets. The results are shown in Table 3. In all three instances, we see an improvement in both precision and F-score, for a comparatively smaller drop in recall. Especially for the 2013 task, the pruned *TEES* system (+0.23pp F-score over *TEES*) matches in performance with the winning 2013 *EVEX* system. Since the *EVEX* system was also based on TEES, it is interesting to note that we have matched these improvements using a different approach. Finally, the pruned *EVEX* system (+0.18pp F-score over the *EVEX* entry) establishes a new top score on the task. Naturally, the magnitude of the F-score improvements is modest, as the top-ranking systems are well optimized to begin with and major improvements have been difficult to achieve regardless of the approach. Note also that a filtering approach such as the one proposed in this paper cannot increase recall because it is unable to produce new events. Our main focus thus is on improving the precision, rather than the recall, aiming to increase the credibility of large-scale event extraction systems in general.

4.2 Tree Organization Before/After Pruning

The tree before and after pruning was visualized, up to depth of 9 using the *Dendroscope* softare (Huson and Scornavacca, 2012). Every intermediate node of the tree was labeled based on the event type that it was mostly associated with and level of purity of that sub-cluster. We noticed before pruning the tree, many sub-clusters associated with different event types are clustered close to each other. How-

Predictions	P	R	F
TEES (ST'11)	61.76	48.78	54.51
Pruned-TEES (ST'11)	62.41	48.75	54.74
TEES (ST'13)	56.32	46.17	50.74
Pruned-TEES (ST'13)	57.20	45.96	50.97
EVEX (ST'13)	58.03	45.44	50.97
Pruned-EVEX (ST'13)	58.85	45.23	51.15

Table 3: Results of official TEES predictions on ST'11 and ST'13 test sets and official EVEX predictions on ST'13 test set before/after pruning. (*P:Precision, R:Recall, F:F-Score*)

ever, after removing sub-clusters that were composed purely of *incorrect* triggers, much better distinctive regions are formed in the tree that are composed of sub-clusters associated with the same event types. Visualizations of the tree are uploaded for further inspections[1].

5 Conclusions

We have investigated hierarchical clustering of event trigger words from the large-scale EVEX event collection, based on their *word2vec* vector space representations and event type distributions. Using the binary cluster tree, we have shown that it is possible to prune even commonly occurring incorrect event triggers and thus achieve a modest improvement over the winning system of the BioNLP 2013 Shared task on GENIA event extraction.

Acknowledgments

We would like to thank Sofie Van Landeghem, Ghent University, for initiating the ideas for this project and her valuable suggestions. We also kindly thank Jari Björne, University of Turku, for his help during the project.

References

Jari Björne and Tapio Salakoski. 2013. TEES 2.1: Automated annotation scheme learning in the BioNLP 2013 Shared Task. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 16–25, Sofia, Bulgaria, August. Association for Computational Linguistics.

Jari Björne, Filip Ginter, and Tapio Salakoski. 2012. University of Turku in the BioNLP'11 Shared Task. *BMC Bioinformatics*, 13(Suppl 11):S4.

Martin Gerner, Farzaneh Sarafraz, Casey M. Bergman, and Goran Nenadic. 2012. Biocontext: an integrated text mining system for large-scale extraction and contextualization of biomolecular events. *Bioinformatics*, 28(16):2154–2161.

Kai Hakala, Sofie Van Landeghem, Tapio Salakoski, Yves Van de Peer, and Filip Ginter. 2013. EVEX in ST'13: Application of a large-scale text mining resource to event extraction and network construction. In *Proceedings of the BioNLP Shared Task 2013 Workshop (BioNLP-ST'13)*, pages 26–34.

Daniel H Huson and Celine Scornavacca. 2012. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Systematic Biology*.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011. Overview of BioNLP Shared Task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 1–6, Portland, Oregon, USA, June. Association for Computational Linguistics.

Haibin Liu, Tom Christiansen, William A Baumgartner Jr., and Karin Verspoor. 2012. Biolemmatizer: a lemmatization tool for morphological processing of biomedical text. *Journal of Biomedical Semantics*, 3(3):on–line, April.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. 2013. Overview of BioNLP Shared Task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1–7, Sofia, Bulgaria, August. Association for Computational Linguistics.

Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski, and Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine (LBM 2013)*, pages 39–43.

Sofie Van Landeghem, Jari Björne, Chih-Hsuan Wei, Kai Hakala, Sampo Pyysalo, Sophia Ananiadou, Hung-Yu Kao, Zhiyong Lu, Tapio Salakoski, Yves Van de Peer, and Filip Ginter. 2013a. Large-scale event extraction from literature with multi-level gene normalization. *PLoS One*, 8(4).

Sofie Van Landeghem, Suwisa Kaewphan, Filip Ginter, and Yves Van de Peer. 2013b. Evaluating large-scale text mining applications beyond the traditional numeric performance measures. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP'13)*, pages 63–71.

[1] http://tinyurl.com/SMBM2014-Paper