UTU: Adapting Biomedical Event Extraction System to Disorder Attribute Detection

Kai Hakala
University of Turku Graduate School (UTUGS), University of Turku, Finland
Dept. of Information Technology, University of Turku, Finland
kahaka@utu.fi

Abstract

In this paper we describe our entry to the SemEval 2015 clinical text analysis task. We participated only in the disorder attribute detection task 2a. Our main goal was to assess how well an information extraction system originally developed for a different task and domain can be utilized in this task. Our system, based on SVM and CRF classifiers, showed promising results, placing 3rd out of 6 participants in this task with performance of 0.857 measured in weighted accuracy, the official evaluation metric.

1 Introduction

SemEval 2015 introduced a new subtask for the clinical text analysis track focusing on disorder mention attribute detection. These attributes describe the relevant information extracted from the textual context of the given disease mention, such as the severity or body location of the disease. The attributes were grouped into 9 separate categories, each with a predefined set of valid attribute classes. The task was defined as a template filling task where the textual cue words for the attributes have to be first identified and then normalized to the correct class. Similar task with slightly different definition has previously been organized as part of the ShARe/CLEF eHealth shared task (Mowery et al., 2014).

Due to time limitations we participated only in the task 2a in which the gold standard disorder mentions were given and only the attribute values had to be predicted. Our main motivation for this years entry was to evaluate the performance of an existing information extraction system, TEES (Björne and Salakoski, 2013), previously developed for a different domain and to assess how easily it can be adapted to a new task.

2 System Description

Turku Event Extraction System (TEES) was originally developed in 2009 for the BioNLP Shared Task on Event Extraction (Kim et al., 2009). This task focused on the extraction of biological processes and interactions between genes and proteins (GGPs) described in biomedical literature. In this task each *event*, i.e. biological process or interaction, is represented by a *trigger* word, which also describes the type of the event, and a set of argument GGP mentions. The argument GGPs may also act in various roles, i.e. each argument is also typed. The participants were thus required to detect these trigger words, their types from a predefined set and the arguments, i.e. the relations between the trigger words and GGPs. Gold standard gene and protein mentions were provided by the organizers and consequently TEES does not include tools for named entity recognition, but presumes these to be given as input data. An example sentence along with the extracted event is illustrated in figure 1.

TEES was the best performing system in the 2009 BioNLP Shared Task as well as in various subtasks in subsequent years (Björne and Salakoski, 2011; Nédellec et al., 2013) showing state-of-the-art performance in biomedical event extraction. Whereas the event extraction task requires the detection of trigger words and argument relations, the disorder attribute detection can be solved by first finding the

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 375–379,
Denver, Colorado, June 4-5, 2015. ©2015 Association for Computational Linguistics

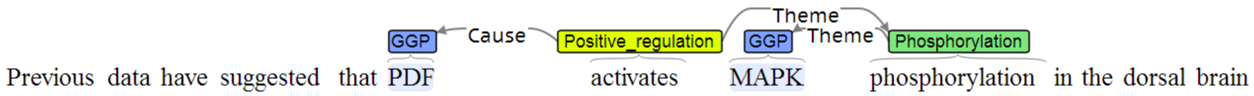
GGP
Cause
Positive_regulation
Theme
GGP
Theme
Phosphorylation
Previous data have suggested that PDF activates MAPK phosphorylation in the dorsal brain

Figure 1: Visualization of an extracted event. In BioNLP Shared Task on Event Extraction the GGP mentions are given and the participants are asked to detect the trigger words, here *activates* and *phosphorylation*, as well as the relation between these entities.

cue words and then relating them to the correct disease mentions, making TEES applicable also for this task.

2.1 Cue Word and Relation Detection

TEES consists of two main processing stages. The first step, called trigger detection, resembles common NER classification task, and classifies each token in the text to either negative class or one of the positive classes, i.e. the predefined trigger types. In this task the trigger detector is used to detect the attribute cue words and their classes.

The second step detects relations between the known named entities and trigger words. This is implemented by generating all plausible entity pairs in a sentence in which case the task becomes a simple classification problem: each pair is classified to either negative class or a positive class resembling the type of the relation.

As trigger and relation detection tasks are both multiclass classification problems, they have been implemented with a multiclass SVM (Tsochantaridis et al., 2004) using the SVM$^{\text{multiclass}}$ software, bundled with TEES. TEES generates a vast amount of classification features from the examined words as well as their context. The relation detection, in particular, relies heavily on syntactic dependencies.

The optimal value for C-parameter is selected independently for each step. However, the independently optimized trigger detection model may not result in the optimal overall system. This is due to the fact that the relation detector is able to discard unwanted triggers, but cannot recover from low trigger detection recall. To overcome this issue, the recall of the trigger detector is artificially increased and the final verdict is made by the relation detector. The amount of overgeneration is selected by evaluating the overall performance of the system.

Whereas TEES relies on graph based data representation with textual entities and the relations between them, the disorder attribute detection task in SemEval 2015 is defined as a slot filling problem. The main issue in the conversion between these two formats is that the default normalization slot values with the corresponding cue defined as *null* cannot be represented in TEES format. Due to this, the default value was decided to be the negative class. In this definition, our system is only aiming to predict the non-default values and if no cue word and a relation between the cue word and disorder entity can be found the default value is preserved. As the slot filling format defines different categories and predefined normalization classes inside these categories, whereas TEES uses a single class for each trigger, the category and normalization classes are concatenated into a single class. E.g. our system is not aware that cue word classes *SV_slight* and *SV_severe* are both normalization values of the severity category, but sees them as independent classes. The relations between cue words and disorder mentions are predicted to only exist or not, i.e. the relations are not typed.

2.2 Body Location Detection

In our evaluation on the development set, the performance of the TEES trigger detector was extremely poor for the body location attributes. This might be due to various reasons. Firstly, whereas the other attribute categories are rather closed sets of expressions, the body locations are named entities. Secondly, TEES does not use any features tailored for the clinical domain and thus generalizes poorly to body location mentions not seen on the training data, resulting in a high precision and low recall system.

As the first attempt to adapt TEES to this task and generalize better for the body locations, we included dictionary features for the trigger detection stage. The used dictionary was composed of the UMLS concepts included in the semantic categories “Body Part, Organ, or Organ Component”, “Body Location or Region", "Body System", "Body Space or Junction", "Body Substance", "Tissue", "Cell" and "Embryonic Structure". These semantic types cover 98.9% of the body locations seen in the training data. For each concept, the preferred term as well as the synonyms were included in the dictionary.

The addition of these features did not improve our performance significantly and thus in the final system, the TEES trigger detector was replaced with a CRF classifier for the body locations. In this approach we used the NERsuite software based on the CRFsuite implementation (Okazaki, 2007). In addition to the standard features such as the word form, lemma, part-of-speech tag and text chunk we incorporated the same dictionary features used in the TEES trigger detector. Moreover, we trained another CRF using the AnatomyTagger software and AnatEM corpus (Pyysalo and Ananiadou, 2013). These two models were stacked, i.e. the predictions from the AnatEM model were given as features for the other classifier.

As the gold standard data includes only attributes related to a disease mention, the annotation is incomplete for NER purposes, and thus using the whole data resulted in poor performance. To prevent this, we trained the body location NER system with only the sentences including at least one annotated body location mention. The development set was filtered in similar fashion for evaluation purposes. The feature set which resulted in the best performance in this evaluation set was used in the final system. This approach boosted the performance on sentences which included at least one annotated body location mention, but the impact on other sentences is hard to assess without complete evaluation data. However, this approach leads to a similar outcome as the aforementioned trigger word overgeneration and shifts the responsibility of removing the excessive body location mentions to the relation detector.

2.3 Disorder and Body Location Normalization

The body location attribute differs from the other categories in that the cue spans were required to be normalized into the corresponding UMLS concepts. As TEES does not include tools for this type of normalization and the normalization was not our main focus in this year's entry, we used a simple tfidf-weighted vector space model. As the first attempt the model was created from the same UMLS concepts used in the body location NER features, but due to high amount of ambiguity this led to poor results. Consequently, we naively generated the model from the gold standard body location annotations and a given entity was then mapped to the UMLS identifier of the most similar entity seen on the training set. If an entity was annotated with various identifiers in different contexts, we used the most frequently occurring identifier.

The entities were predicted to be "CUI-less" if the most similar gold standard entity was annotated as such or if the maximum cosine similarity was zero. Thus in this naive approach there was no need for more complex "CUI-less" value identification as is necessary in our previously suggested normalization method (Kaewphan et al., 2014).

The disorder mention normalization was not part of the original slot filling task, but was later on added to the task definition. For simplicity we used the same naive method as with the body location entities.

3 Results

We submitted three separate runs to the final evaluation. Runs 1 and 2 used the same approach, but run 2 includes a last-minute bug fix which we were not able to thoroughly test. This bug caused some of the attribute mentions to be duplicated during the conversion between SemEval and TEES data formats, misleading the system. These runs use the method described in this paper, but the system was only allowed to predict one value for each slot. This was forced by only selecting the value with highest classification confidence for the relation detection; the confidence of the trigger word detection was ignored. In run 3 we allowed the system to predict multiple body location values for each disorder mention. This is beneficial in statements such as *"Osteophytes are seen along the medial tibial plateau as well as the superior aspect of the patella"* where both body locations *tibial plateau* and *patella* are related to the same disorder mention *Osteophytes*. The results for these runs are shown in table 1 along with the best runs from the other participated groups.

Our best performance was obtained from the run

Team	WA	A
UTH-CCB	0.886	0.943
ezDI	0.880	0.934
UTU run3	0.857	0.945
UTU run2	0.855	0.944
UTU run1	0.846	0.939
UWM	0.818	0.859
TeamHCMUS	0.576	0.195
UtahPOET	0.446	0.744

Table 1: Official test set results for our 3 submissions and the other 5 participating teams. Only the best runs measured in weighted accuracy are shown for other teams. WA = weighted accuracy, A = non-weighted accuracy.

3 with weighted accuracy of 0.857, resulting in the third best performing system in the task. Measured on the non-weighted accuracy which was not the main evaluation metric, but still included in the official results, we achieved score 0.945, the second best performance in the task.

Runs 1 and 2 which did not allow multiple body locations to be predicted performed slightly worse, run 2 achieving weighted accuracy of 0.855. This difference between runs 2 and 3 is solely caused by the body location category in which the difference between these two runs is 1.1pp. The category-wise performance is shown in table 2.

The comparison of our results to the best performing system by team UTH-CCB reveals that our system performs consistently weaker in every category. Worth noticing is that our naive normalization approach is not affecting our performance dramatically, showing weighted accuracy of 0.827 in disorder normalization category (CUI), where as UTH-CCB system achieved score of 0.854. As the gold standard disorder mentions were given in this task, this score is only measuring the normalization performance.

Our submitted runs were all trained with the combination of training and development data sets. The overall results on development and test sets are fairly similar showing that the system is not overfitting to the development data. On the other hand it seems that combining the training and development sets for the final models does not improve the performance significantly, although we cannot confirm this speculation before the gold standard annotation for the test data is released. As an exception to this is our normalization method, which greatly benefits from the added training data as can be seen from the +5.5pp improvement in the CUI category. This shows that the naive approach does not generalize well and is applicable only when the training data covers most of the disorder mentions seen in the test data.

4 Discussion and Future Work

The current implementation of TEES induces some limitations for this task. Firstly, the current data format used in TEES does not allow the representation of discontiguous entities, which are not common in various other tasks. In this submission we thus represented the discontiguous disorder entities with a single span during the cue word and relation detection. As the discontiguous entities are much less frequent in the attribute entities, we discarded them completely. As a future work we would like to allow TEES to support this type of entities. This will require not only altering the used data format, but also modifying the feature extraction process to be able to fully express the characteristics of these entities.

Secondly, TEES uses micro-averaged F-score of positive classes as the internal evaluation metric for parameter optimization, which may be suboptimal for tasks evaluated in different metrics. Due to this, we plan to modify TEES to accept various user-defined evaluation metrics.

To improve our performance in this task specifically, we need to first perform a detailed error analysis. This might reveal for instance whether some domain specific features could improve the accuracy of our system.

5 Conclusions

We have demonstrated that an information extraction system originally developed for scientific literature can be easily adapted to the clinical domain. The described system shows competitive performance being the third best system in the disorder attribute slot filling task. We have also discussed some of the limitations of the system and suggested multiple future improvements for better suitability to new task definitions and domains.

Team	WA	A	BL	CUI	CND	COU	GEN	NEG	SEV	SUB	UNC
UTH-CCB	0.886	0.943	0.862	0.854	0.903	0.887	0.911	0.975	0.936	0.975	0.911
Run3	0.857	0.945	0.825	0.827	0.823	0.798	0.888	0.970	0.915	0.920	0.853
Run2	0.855	0.944	0.814	0.827	0.823	0.798	0.888	0.970	0.915	0.920	0.853
Run3 devel	0.830	0.933	0.798	0.772	0.862	0.848	0.864	0.941	0.940	0.920	0.872

Table 2: Performance of our system in each attribute category compared to the best performing system. *Run 3 devel* shows our best results for the development set evaluated with the evaluation tool provided by the organizers.

Acknowledgements

Computational resources were provided by CSC — IT Center for Science Ltd, Espoo, Finland.

References

Jari Björne and Tapio Salakoski. 2011. Generalizing biomedical event extraction. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 183–191, June.

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

Suwisa Kaewphan, Kai Hakala, and Filip Ginter. 2014. UTU: Disease mention recognition and normalization with CRFs and vector space representations. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 807–811, August.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1–9, June.

Danielle L. Mowery, Sumithra Velupillai, Brett R. South, Lee Christensen, David Martinez, Liadh Kelly, Lorraine Goeuriot, Noemie Elhadad, Sameer Pradhan, Guergana Savova, and Wendy Chapman. 2014. Task 2: ShARe/CLEF eHealth Evaluation Lab 2014. In *Proceedings of CLEF 2014*, September.

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, August.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of conditional random fields (CRFs). http://www.chokkan.org/software/crfsuite/.

Sampo Pyysalo and Sophia Ananiadou. 2013. Anatomical entity mention recognition at literature scale. *Bioinformatics*.

Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. 2004. Support vector machine learning for interdependent and structured output spaces. In *International Conference on Machine Learning (ICML)*, page 104.