NTU-1 at SemEval-2017 Task 12: Detection and classification of temporal events in clinical data with supervised domain adaptation

Po-Yu Huang^{*}, Hen-Hsen Huang[†], Yu-Wun Wang[†], Ching Huang[†], Hsin-Hsi Chen[†]

*School of Medicine

Taipei Medical University Taipei, Taiwan [†] Department of Computer Science and Information Engineering National Taiwan University Taipei, Taiwan

Abstract

This study proposes a system to automatically analyze clinical temporal events in a fine-grained level in SemEval-2017. Support vector machine (SVM) and conditional random field (CRF) were implemented in our system for different subtasks, including detecting clinical relevant events and time expression, determining their attributes, and identifying their relations with each other within the document. Domain adaptation was the main challenge this year. Unified Medical Language System was consulted to generalize events specific to each domain. The results showed our system's capability of domain adaptation.

1 Introduction

This study proposes a system to participate in the Clinical TempEval 2017 shared task, which focuses on the detection and classification of temporal events in clinical data. To better utilize the information in clinical data, temporal event extraction is fundamental in previous researches (Bethard et al., 2016). Unlike previous studies, the training and the test data are in different domains this year. The task is further separated into two phases: unsupervised domain adaption and supervised domain adaption. We took part in the supervised domain adaption where data of 591 records of colon cancer patients and 30 records of brain cancer patients from Mayo clinic were given. Based on the THYME corpus (Styler IV et al., 2014), we propose a framework that automatically analyzes clinical temporal events in a finegrained level. Our framework identifies temporal

events in unstructured text and further labels every event with its attributes.

The task consists of three major subtasks. The first one is to detect clinical relevant events and time expressions in a given medical record. From the unstructured text, both the spans of time expressions and the spans of event mentions are identified.

The second subtask is analyzing the attributes of time expressions and event mentions. A time expression will be classified into one of six types: DATE, TIME, DURATION, QUANTIFIER, PREPOSTEXP, and SET. An event mention contains four properties such as type of an event, polarity, degree and modality. Our model labels these four properties to every event mention.

The final subtask is to determine two kinds of relations within the text. DocTimeRel is the relation between the document creation time and an event mention. Four types of DocTimeRel including BEFORE, AFTER, OVERLAP, and BE-FORE-OVERLAP are annotated in THYME. In addition to DocTimeRel, our model also recognizes the narrative container of an event mention called TLINK in this task. There are five types of TLINK, including BEFORE, CONTAIN, OVER-LAP, BEGINS-ON, and ENDS-ON.

The outcomes of our system are not only the clinical temporal events, but also their detailed properties and their temporal relations with other events. The results of our framework can be further used to discover relationships between illnesses, symptoms, medications, and procedures.

2 Methods

Our system contains five modules: the first one identifies the span and the type of each event mention; the second one determines the other remaining attributes of an event mention; the third one identifies the span and the type of each time expression; the fourth one determines the TLINK between each pair of event mention and time expression in the same sentence; the fifth one determines the TLINK between each pair of event mentions in the same sentence.

2.1 Preprocessing

We ran Stanford CoreNLP toolkit (Manning et al., 2014) on all the clinical data. This toolkit generated POS and NER for each word, and dependencies for each sentence in the clinical data. A dictionary was built based on Unified Medical Language System (UMLS) (Bodenreider, 2004) with five categories of different genre of medical related words, including DIAGNOSIS, EXAMINE, MEDICINE, POSITION and SURGERY. All of these were utilized in the following steps.

2.2 Event Mention Identification

In this module, we tried to identify the span denoting an event and its type. There are three types of event mentions, including ASPECTUAL, EV-IDENTIAL, and N/A. ASPECTUAL event mentions often turn out to be verbs indicating something would happen later in the timeline, like "reoccur", "continue", etc. EVIDENTIAL events are usually verbs like "show", "reveal", and "confirm", which show how doctors come to identify and learn about other events. N/A events are mostly composed of medical related words like "nausea", "surgery", and "operate".

We built a four-way linear SVM classifier using scikit-learn (Pedregosa et al., 2011) to classify a word into ASPECTUAL, EVIDENTIAL, N/A, or non-event. In other words, span identification and type classification are done simultaneously. Features for our SVM classifier are listed as follows:

Lexical Feature: n-gram of nearby words, and character n-gram within the target word

POS Feature: POS n-gram of nearby words

Named Entity Type: type of named entities identified by NER

Orthographic Feature: orthographic n-gram of nearby words obtained by substituting all uppercase letters, lowercase letters, and digits with 'A', 'a', and '0'.

Structural Features: 1) position of the target word divided by the sentence length, and 2) the

length of the path from the target word to the root in the dependency-parsing tree.

UMLS Category: the category of the target word based on UMLS.

Since training set and test set come from two different domains, i.e., colon cancers versus brain cancers, there may be some medical terms in test set but not appearing in training set. In this study, UMLS dictionary was consulted to cluster the medical terms into the five categories in order to deal with domain adaptation problem and reduce sparseness. For example, terms specific to colon cancer "right hemicolectomy" and "rectum" would be transformed into "SURGERY" and "POSITION", respectively, while building n-gram features.

2.3 Event Attribute Identification

Besides modality, polarity, and degree, we include DocTimeRel from the relation subtask here because it is also an attribute along with an event mention. We trained a multi-class linear SVM classifier for each of the four attributes. The features described in Section 2.2 were used. In additions, we introduced time related features for DocTimeRel, including tense of verbs within the same sentence, n-gram and POS n-gram of time related terms within the same sentence, and the relative position of the time related terms within the same sentence.

2.4 Time Expression Identification

Time expression identification is different from event mention identification because time expression is less affected by the change of domain. However, its spans are more diverse than those in event mention. For instance, a SET time expression is usually composed of multiple words like "three times a week". By contrast, a PRE-POSTEXP is mostly just one word only, like "preoperative". To deal with the issue of diverse spans, we used CRF¹ to develop this module because of its ability in sequence labeling. Similarly to Section 2.2, we also determined the span and the type simultaneously.

Besides those features (UMLS Category excluded) used in Section 2.2, we added some other features, including the existence of pre-post related characters ("pre", "post", "peri", and "intra"), the existence of a number, and whether there is a

¹ http://sklearn-crfsuite.readthedocs.io/en/latest/index.html

duration condition in the same sentence ("for", "since", "through", "until", and "in").

2.5 TLINKs between Event Mentions and Time Expressions

TLINK is determined between event mentions, and between event mentions and time expressions. TLINKs are mostly linked within the same sentence, thus we focused on identifying TLINKs within sentences.

Two multi-class SVM classifiers were built for five subtypes in TLINK: the first one was trained to identify TLINKs given a pair of time expression and event mention, which we called "TE classifier", and the second one was trained to identify TLINKS given a pair of event mentions, which we called "EE classifier".

Features we used are shown as follows:

Features for both classifiers: types, attributes, tokens and POS of the pair of mentions, punctuations between the pair of mentions, tense of the nearest verbs, and dependency path between the pair of mentions.

Features only for EE classifier: if exists a time expression which is linked to both event mention by the TE classifier, types of the two TLINKs were considered as features

3 Results

We used the clinical data provided in the supervised domain adaption, which contained 591 records of colon cancer patients and 30 records of brain cancer patients, to train our system. Table 1 shows the results of event mentions, time expression and relations, where F1 stands for F1 score, P stands for precision, and R stands for recall.

Brain cancer	F1	Р	R
ES	0.73	0.62	0.87
ES: All attributes	0.41	0.35	0.50
ES: Modality	0.63	0.54	0.75
ES: Degree	0.72	0.62	0.86
ES: Polarity	0.70	0.60	0.84
ES: Type	0.70	0.60	0.85
DocTimeRel	0.49	0.42	0.59
TS	0.58	0.58	0.58
TS: Type	0.54	0.54	0.54
TLINK	0.26	0.20	0.37

Table 1: Results of event spans (ES), time spans (TS), TLINK, and their attributes tested on brain cancer patients.

To compare our system's performance while switching domain, we also provide Table 2 of the results for all three subtasks where training and testing data are all colon cancer patients.

Colon cancer	F1	Р	R
ES	0.86	0.84	0.89
ES: All attributes	0.57	0.55	0.58
ES: Modality	0.78	0.76	0.81
ES: Degree	0.86	0.84	0.88
ES: Polarity	0.83	0.81	0.86
ES: Type	0.83	0.81	0.86
DocTimeRel	0.65	0.63	0.67
TS	0.75	0.83	0.68
TS: Type	0.73	0.80	0.66
TLINK	0.39	0.38	0.39
• •			

Table 2: Results of all subtasks tested on colon
cancer patients.

4 Discussion

The F1 scores of event mentions in brain cancer patients are lower than in colon cancer patients. It is mostly contributed by the decrease in precision. Without the ground truths, we can only assume that our system still learned some domain-specific features to tell event apart from non-event under the replacement of words with classes according to UMLS. The scores of TLINK have the same problem as event mentions.

Interestingly, the performance of time expression, which we thought to be free from the challenge of domain adaptation, decreases drastically in all three scores. It is certain that some domain specific features played big roles in our system. However, without the ground truths, it is difficult to identify the problem.

CRF and SVM were both experimented for time expressions and shown in Table 3. With the same features, CRF performed better than SVM in F1 score and precision. The results show that CRF has a better performance in sequence labeling. Advanced deep learning model including convolution neural network (CNN) and recurrent neural network (RNN) will be explored in the future due to their advantages in sequence labeling.

Time	F1	Р	R
SVM	0.69	0.67	0.70
CRF	0.73	0.80	0.66

Table 3: results of time expression identification with two different settings

Best vs. Us	F1	Р	R
ES	0.76 (0.73)	0.69 (0.62)	0.85 (0.87)
ES: All attributes	0.52 (0.41)	0.47 (0.35)	0.58 (0.50)
ES: Modality	0.69 (0.63)	0.63 (0.54)	0.78 (0.75)
ES: Degree	0.75 (0.72)	0.68 (0.62)	0.84 (0.86)
ES: Polarity	0.75 (0.70)	0.68 (0.60)	0.83 (0.84)
ES: Type	0.75 (0.70)	0.68 (0.60)	0.83 (0.85)
DocTimeRel	0.59 (0.49)	0.53 (0.42)	0.66 (0.59)
TS	0.58 (0.58)	0.51 (0.58)	0.67 (0.58)
TS: Type	0.55 (0.54)	0.49 (0.54)	0.64 (0.54)
TLINK	0.32 (0.26)	0.25 (0.20)	0.43 (0.37)

Table 4: Comparison with the best results in F1 of the other runs in this shared task, where our results are listed inside the brackets.

Compared to the best results of other runs in this shared task, which is shown in Table 4, Doc-TimeRel is the most poorly predicted attribute by our system. DocTimeRel is the relationship between an event and its document creation time. However, the features we chose for this subtask were all confined to one sentence. Adding features capturing the time representations within neighboring sentences, within the section, or even within whole document should increase the performance.

TLINK is another attributes that our system performed notably worse than the first place. This is possibly due to the chain effect where TLINK was determined based on event mentions and time expressions that were already with worse performance. Once ruling out this possibility, we can then focus on how to improve our TLINK module.

5 Conclusion

In this paper, we proposed a system to participate in the Clinical TempEval 2017 shared task. Our system not only identified the clinical temporal events, but also their detailed properties and their temporal relations with other events. It can also take on the challenge of domain adaptation where only a few data from targeted domain was given while the other data were from different domain. Our system adopted SVM and CRF for different subtasks. The results were in the third place in supervised domain adaptation.

In future works, we will focus on the increasing the performance in DocTimeRel and explore deep learning algorithms.

References

- Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. (2016). Semeval-2016 task 12: Clinical tempeval. *Proceedings of SemEval*:1052-1062.
- Oliver Bodenreider. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res, 32*(Database issue):D267-270. doi:10.1093/nar/gkh061
- Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. (2014). *The stanford corenlp natural language processing toolkit*. In *Proceedings of the ACL (System Demonstrations)*, pages 55-60.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel. Mathieu Blondel. Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. (2011). Scikit-learn: Machine learning in Python. Journal Machine Learning of Research, 12(Oct):2825-2830.
- William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, and Guergana Savova. (2014). Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143-154.