UTHealth at SemEval-2016 Task 12: an End-to-End System for Temporal Information Extraction from Clinical Notes

Hee-Jin Lee, Yaoyun Zhang, Jun Xu, Sungrim Moon, Jingqi Wang, Yonghui Wu, and Hua Xu
University of Texas Health Science Center at Houston
Houston, TX, USA
hee.jin.lee@uth.tmc.edu, {firstname.lastname}@uth.tmc.edu

Abstract

The 2016 Clinical TempEval challenge addresses temporal information extraction from clinical notes. The challenge is composed of six sub-tasks, each of which is to identify: (1) event mention spans, (2) time expression spans, (3) event attributes, (4) time attributes, (5) events' temporal relations to the document creation times (DocTimeRel), and (6) narrative container relations among events and times. In this article, we present an end-to-end system that addresses all six sub-tasks. Our system achieved the best performance for all six sub-tasks when plain texts were given as input. It also performed best for narrative container relation identification when gold standard event/time annotations were given.

1 Introduction

Temporality is crucial in understanding the course of clinical events from a patient's electronic health records. Since a large part of the information on temporality resides in narrative clinical notes, automatic extraction of temporal information from clinical notes using natural language processing (NLP) techniques has received much attention. Over the years, research community challenges on clinical temporal information extraction have been organized; i.e., the 2012 Informatics for Integrating Biology and the Bedside (i2b2) challenge (Sun et al., 2013), the 2013/2014 CLEF/ShARe challenge (Mowery et al., 2014), and the 2015 Clinical TempEval challenge (Savova et al., 2015). These challenges provide annotated corpora on temporal entities and relations, which facilitate comparisons of multiple systems and expediate the development of clinical temporal information extraction methodologies.

The 2016 Clinical TempEval challenge is the most recent community challenge that addresses temporal information extraction from clinical notes. Following the 2015 Clinical TempEval challenge, the 2016 challenge consists of six sub-tasks, each of which is to identify: (1) spans of event mentions, (2) spans of time expressions, (3) attributes of events, (4) attribute of times, (5) events' temporal relations to the document creation times (DocTimeRel), and (6) narrative container relations among events and times (TLINK:Contains). 440 annotated clinical notes from Mayo Clinic, or the THYME corpus (Styler IV et al., 2014), were provided as the training data set, and 153 plain text clinical notes were provided as the test set. The participating systems were evaluated through two phases. In phase 1, the systems were evaluated on their results for all six sub-tasks given plain texts as inputs. In phase 2, system predictions on DocTimeRel and TLINK:Contains were evaluated given the gold-standard event annotations (EVENT) and time annotations (TIMEX3).

In this article, we describe a comprehensive system that addresses all six sub-tasks. We designed the system by adapting state-of-the-art techniques from previous work on named entity recognition (Tang et al., 2013a; Jiang et al., 2011) and temporal relation identification (Tang et al., 2013b; Lin et al., 2015) in the medical domain. Our end-to-end system achieved top performance for all six sub-tasks in the phase 1 and the TLINK:Contains identification task in the phase 2 stages of the challenge.

Proceedings of SemEval-2016, pages 1183–1188,
San Diego, California, June 16-17, 2016.

2 Methods

Our temporal information extraction system consists of four modules: the first module identifies the spans of event mentions and time expressions along with their types; the second module identifies attributes of events and times; the third module predicts DocTimeRel; and the last module identifies TLINK:Contains among events and times. The output results from previous modules are utilized by the latter modules. We describe those modules in detail in the following sections.

Please note that we utilized the following tools to construct our system: 1) CLAMP toolkit (`http://clinicalnlptool.com/index.php`) for tokenization, 2) OpenNLP toolkit (`http://opennlp.sourceforge.net/`) for Part-Of-Speech (POS) tagging and consitituency parsing, and 3) ClearNLP toolkit (`https://code.google.com/p/clearnlp/`) for dependency parsing. We utilized wrappers provided by cTAKES (Savova et al., 2010).

2.1 Event mentions and temporal expressions recognition

As the first step, our system identifies the spans of event mentions and time expressions along with their types.

According to our observations of the corpus, different types of event mentions and time expressions may show characteristics different from one another. For instance, events with EVIDENTIAL type are usually represented with verbs such as 'showed', 'reported', 'confirms', in contrast to the events with N/A type that are usually represented with medical terms such as 'nausea', 'chemotherapy' or 'colonoscopy'. Similarly, times with DATE type appear more often with the preposition 'on', while times with DURATION type appear more often with 'during' or 'since'. Such variations in the characteristics may limit the system's performance, if one tries to identify event mentions or time expressions of all types at once and then identify their types. Therefore, our system identifies the spans of events and times as well as their types simultaneously.

An HMM-SVM sequence tagger (Joachims et al., 2009) is employed to tag each token in the clinical notes as either O (outside of an event mention), B-*type* (beginning of an event mention of *type*), or I-*type* (inside of an event mention of *type*), where *type* can be any of the three event types defined by the Clinical TempEval challenge (i.e, N/A, ASPECTUAL, and EVIDENTIAL). Another HMM-SVM tagger is used in a similar manner to identify spans and types of time expressions.

We use various features that have been successfully used for many entity recognition tasks in the clinical domain (Tang et al., 2013b; Lin et al., 2015). In addition, we incorporate the results of SUTime (Stanford temporal tagger) (Chang and Manning, 2012) into our system as a feature. SUTime is a rule-based tagger that identifies time expressions as defined by the TimeML (Mani and Pustejovsky, 2004). The features used are as follows:

Lexical features: n-gram (uni-, bi-, and tri-) of nearby words (window size of +/- 2), character n-gram (bi- and tri-) of each word, prefix and suffix of each word (up to three characters), and orthographic forms of each word (obtained by normalizing numbers, uppercase letters, and lowercase letters to '#', 'A', and 'a', respectively, and by regular expression matching)

Syntactic features: POS n-gram (uni-, bi-, and tri-) of nearby words (window size of +/- 2)

Discourse level features: sentence length, sentence type (e.g., whether the sentence ends with a colon or starts with an enumeration mark such as '1.'), and section information

Word representation features: features derived from Brown clustering (Brown et al., 1992), random indexing (Lund and Burgess, 1996) and word embedding (Tang et al., 2014) (trained on MiPACQ (Albright et al., 2013) and MIMIC II (Saeed et al., 2011) corpora)

Features from external resources: dictionary matching results using customized dictionaries of medical/temporal terms, and the temporal expression prediction results from SUTime (TIMEX3 only).

2.2 Event attribute identification

Given spans and types of event mentions, our system further identifies three attributes of the events,

i.e., *modality*, *degree*, and *polarity*. We trained three SVM classifiers for each of the three attributes using LIBLINEAR SVM package (Fan et al., 2008). We used features similar to those described in Section 2.1, where the features are extracted from a window size of +/- 5 tokens around each event mention. Additionally, we used the attribute-specific features (described below) for event attribute identification.

Attribute-specific features: existence of conditionality/possibility keywords (e.g., 'if', 'unless', 'could', and 'likely') in the window size of +/-5 tokens

2.3 DocTimeRel identification

Our system identifies DocTimeRel of each event mention in a manner similar to which it identifies event attributes. An SVM classifier was trained using the LIBLINEAR package, where the features are extracted from the window of +/- 5 tokens around each event mention. In addition to the set of features similar to the ones described in Section 2.1, the following features are used:

DocTimeRel-specific features: tense information of the verbs in the same sentence, event attributes, and information on time expressions in the same sentence (token/POS of time expressions before/after the event mention, token/POS of words between the closest time expression and the event mention)

2.4 TLINK:Contains identification

We divide the task of narrative container relation identification into six sub-problems based on two criteria: (1) whether the target narrative container relation is between two events or between an event and a time and (2) whether the two event/time mentions are within one sentence, within two adjacent sentences, or across more than two sentences. For each sub-problem, we trained an SVM classifier that identifies whether an ordered pair of two events/times (or a *candidate pair*) forms a TLINK of Contains type, using the LINLINEAR SVM package.

Before training the classifiers, we apply the following steps in order to take into account the data distribution characteristics. First, in the gold standard dataset, a large number of implicit temporal relations are left unannotated intentionally. Since providing implicit relations as negative instances to the SVM learners may harm the learning process, we extended the gold standard set of TLINK:Contains to its transitive closure, and used the extended set as the positive instances for training. The transitive closure was generated by applying Floyd-Warshall algorithm (Floyd, 1962) on the gold standard TLINK set based on the transitivity of the TLINK:Contains relation (i.e, A contains B $\wedge$ B contains C $\rightarrow$ A contains C).

Second, since any two events/times can be a candidate pair to train a classifier, the number of candidate pairs becomes huge with small portion of positive instances among them. This may not be ideal for training a classifier. In order to reduce the number of prospective negative instances, we filtered out some of the candidate pairs that are highly unlikey to form a TLINK:Contains relation based on the THYME corpus annotation guideline[1]. We removed a candidate pair either 1) when the two event/time mentions are not in the same section, or 2) when one event has ACTUAL modality while the other has HYPOTHETICAL modality, or 3) when one event has BEFORE DocTimRel while the other has AFTER DocTimeRel. For candidate pairs whose event/time mentions are across more than two sentences, we further filtered out the pairs based on heuristic rules, in order to keep only the candidate pairs that are higly likely to form a TLINK:Contains relation. We kept a candidate pair only when an event/time among the two events/times is mentioned in a section header that includes the keywords 'history' or 'evaluation' or in a section header that ends with a time expression.

We also applied cost-sensitive learning in order to counterbalance the effect of dominating number of negative instances. To each class, we assigned weight that is inversely proportional to the class frequency, adjusting the penalty factor in SVM training (Ben-Hur and Weston, 2009). For instance, if there were 20 positive pairs among 100 candidate pairs, we would assign the weight 5 (100/20) to the positive class and the weight 1.25 (100/80) to the negative class.

[1] http://clear.colorado.edu/compsem/documents/THYME\%20Guidelines.pdf

sub-task	P	R	F
ES (t)	0.915	0.891	0.903
TS (t)	0.840	0.758	0.795
ES (s)	**0.915**	**0.891**	**0.903**
TS (s)	0.836	0.757	**0.795**
ES (m)	0.887	0.846	0.874
TS (m)	0.779	0.539	0.637

Table 1: Test set results on EVENT span (ES) and TIMEX3 span (TS) identification. Bold faces signify the cases where our system showed the top performance.

The features used for the six classifiers are as follows. Note that an event mention was expanded to its covering noun phrase before the feature extraction:

Common features: event/time attributes, token and POS features on event/time mentions (as provided by cTAKES), punctuation between event/time mentions, other event/time mentions within the same sentence as the two event/time mentions, number of other event/time mentions between the two event/time mentions, tense of the verbs in the same sentence, section information, sentence type (the same as in Section 2.1), and word embedding representations of the head words of event/time mentions

Feaures for single-sentence cases: dependency path linking the two event/time mentions (as provided by cTAKES)

Features for multi-sentence cases: line distance between the two event/time mentions, and tokens that are common to the two event/time mentions

3 Results

In this section, we present our system's performance on test set along with the top and the median results from the challenge. Table 1, 2, and 3 show the results on event/time span identification, event/time attribute identification, and DocTImeRel and TLINK:Contains identification, respectively. In the tables, (t) and (m) stand for the top and median results of the 2016 challenge, while (s) stands for our system's results. Our system showed top F1 scores for event/time span, for event/time attribute, for DocTimeRel (phase 1 only), and for TLINLK:Contains identification.

sub-task	P	R	F	A
TA:type (t)	0.815	0.735	0.772	0.989
EA:type (t)	0.894	0.870	0.882	0.977
EA:modality (t)	0.866	0.843	0.855	0.947
EA:degree (t)	0.911	0.887	0.899	0.997
EA:polarity (t)	0.900	0.875	0.887	0.983
TA:type (s)	0.812	**0.735**	**0.772**	0.971
EA:type (s)	**0.894**	**0.870**	**0.882**	**0.977**
EA:modality (s)	**0.866**	**0.843**	**0.855**	**0.947**
EA:degree (s)	**0.911**	**0.887**	**0.899**	0.996
EA:polarity (s)	**0.900**	**0.875**	**0.887**	0.982
TA:type (m)	0.755	0.499	0.618	0.970
EA:type (m)	0.854	0.813	0.844	0.967
EA:modality (m)	0.830	0.780	0.810	0.930
EA:degree (m)	0.882	0.838	0.869	0.995
EA:polarity (m)	0.868	0.900	0.875	0.887

Table 2: Test set results on EVENT attribute (EA) and TIMEX3 attribute (TA) identification. Bold faces signify the cases where our system showed the top performance.

(a)

sub-task	P	R	F
DR (t)	0.766	0.746	0.756
CR (t)	0.531	0.471	0.479
DR (s)	**0.766**	**0.746**	**0.756**
CR (s)	0.488	**0.471**	**0.479**
DR (m)	0.655	0.624	0.639
CR (m)	0.491	0.235	0.318

(b)

sub-task	P	R	F	Acc.
DR (t)	-	-	-	0.843
CR (t)	0.823	0.564	0.573	-
DR (s)	-	-	-	0.835
CR (s)	0.588	0.559	**0.573**	-
DR (m)	-	-	-	0.724
CR (m)	0.589	0.345	0.449	-

Table 3: Test set results on DocTimeRel (DR) and TLINK:Contains (CR) identification. (a) phase 1 results. (b) phase 2 results. Bold faces signify the cases where our system showed the top performance.

DocTimeRel value	# EVENTs (%)	Acc.
AFTER	2073 (10.9%)	81.3%
OVERLAP	8983 (47.3%)	90.1%
BEFORE	6984 (36.8%)	79.9%
BEFORE/OVERLAP	952 (5.0%)	51.5%

Table 4: DocTimeRel identification accuracy on each DocTimeRel value. # EVENTs represents the number of event annotations from the gold standard with the specified DocTimeRel value in the test set.

4 Conclusion and discussion

In this article, we describe a system that shows the top performance in the 2016 Clinical TempEval challenge. We adapted the state-of-the-art techniques for entity recognition and temporal relation identification in the clinical domain, and show that those techniques are effective for the Clinical TempEval challenge as well.

For time expression identification, we found some error cases in which the system's prediction differs with the gold standard annotation only on the inclusion or exclusion of a preposition. For example, while a DURATION type time is annotated for the phrase "for the past 40 years" in the gold set, our system predicted a DURATION for the phrase "the past 40 years" omitting the preposition 'for' from the gold standard annotation.

Table 4 shows the DocTimeRel identification accuracy on each DocTimeRel value. Accuracy on the value OVERLAP is the highest, which might come from the abundance of the training data. Surprisingly, the classifier worked better for the value AFTER than the value BEFORE, even though there were three times more events with BEFORE DocTimeRel than those with AFTER. We conjecture that explicit keywords that indicate the future tense such as "will" and "potential" played key roles in identifying AFTER DocTimeRel.

Table 5 shows the 10-fold cross validation results of the six classifiers for TLINK:Contains identification. Temporal relations between an event and a time were predicted more accurately than the relations between two events. Classifiers for pairs across more than two sentences showed the best F1 scores, due to the heuristic filtering steps in which we kept only the candidate pairs that are highly likely to form a narrative container relation.

sub-problem	F
EVENT-EVENT-1	66.9%
EVENT-EVENT-2	69.1%
EVENT-EVENT-3	76.2%
EVENT-TIMEX3-1	79.9%
EVENT-TIMEX3-2	76.3%
EVENT-TIMEX3-3	84.3%

Table 5: F1 scores of the six classifiers for TLINK:Contains identification (10-fold cross validation on the training set). EVENT-EVENT and EVENT-TIMEX represent the sub-problems regarding the candidate pairs between two events, and the sub-problems regarding the pairs between an event and a time, respectively. The suffixes '-1', '-2' and '-3' indicate that the pairs should be within one sentence, within two adjacent sentences, and across more than two sentences, respectively.

We plan to further improve our system to show higher performance based on the observations above.

Acknowledgments

We would like to thank the Mayo Clinic for permission to use the THYME corpus. This study was supported in part by NIGMS grant 1 R01 GM103859.

References

Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, William F Styler, Colin Warner, Jena D Hwang, Jinho D Choi, Dmitriy Dligach, Rodney D Nielsen, James Martin, Wayne Ward, Martha Palmer, and Guergana K Savova. 2013. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*, 20(5).

Asa Ben-Hur and Jason Weston. 2009. A User's Guide to Support Vector Machines. In *Data Mining Techniques for the Life Sciences*, pages 223–239. Humana Press, Totowa, NJ.

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. *Computational Linguistics*, 18(4):467–479.

Angel X Chang and Christopher D Manning. 2012. SUTIME: A Library for Recognizing and Normalizing Time Expressions. In *Proceedings of 8th International Conference on Language Resources and Evaluation LREC (LREC 2012)*.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Li-

brary for Large Linear Classification. *The Journal of Machine Learning Research*, 9:1871–1874.

Robert W Floyd. 1962. Algorithm 97: Shortest path. *Communications of the ACM*, 5(6):345.

Min Jiang, Yukun Chen, Mei Liu, S Trent Rosenbloom, Subramani Mani, Joshua C Denny, and Hua Xu. 2011. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association*, 18(5):601–606.

Thorsten Joachims, Thomas Finley, and Chun-Nam John Yu. 2009. Cutting-Plane Training of Structural SVMs. *Machine Learning Journal*, 77(1):27–59.

Chen Lin, Dmitriy Dligach, Timothy A Miller, Steven Bethard, and Guergana K Savova. 2015. Multilayered temporal modeling for the clinical domain. *Journal of the American Medical Informatics Association*, pages ocv113–9.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208.

Inderjeet Mani and James Pustejovsky. 2004. Temporal discourse models for narrative structure. In *Proceedings of ACL Workshop on Discourse Annotation*, pages 57–64. Association for Computational Linguistics.

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

Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. 2011. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. *Critical care medicine*, 39(5):952–960.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Guergana Savova, Marc Verhagen, Steven Bethard, Leon Derczynski, and James Pustejovsky. 2015. Semeval-2015 task 6: Clinical tempeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation*.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics*, 2(0):143–154.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Buzhou Tang, Hongxin Cao, Yonghui Wu, Min Jiang, and Hua Xu. 2013a. Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. *BMC Medical Informatics and Decision Making*, 13(Suppl 1):S1.

Buzhou Tang, Yonghui Wu, Min Jiang, Yukun Chen, Joshua C Denny, and Hua Xu. 2013b. A hybrid system for temporal information extraction from clinical text. *Journal of the American Medical Informatics Association*, 20(5):828–835.

Buzhou Tang, Hongxin Cao, Xiaolong Wang, Qingcai Chen, and Hua Xu. 2014. Evaluating Word Representation Features in Biomedical Named Entity Recognition Tasks. *BioMed Research International*, 2014(2):1–6.