Hitachi at SemEval-2017 Task 12: System for temporal information extraction from clinical notes

Sarath P R[1], **Manikandan R**[1], **Yoshiki Niwa**[2]
[1]Research and Development Center, Hitachi India Pvt Ltd, India
[2]Hitachi Ltd, Center for Exploratory Research, Japan
{sarath,manikandan}@hitachi.co.in
yoshiki.niwa.tx@hitachi.com

Abstract

This paper describes the system developed for the task of temporal information extraction from clinical narratives in the context of the 2017 Clinical TempEval challenge. Clinical TempEval 2017 addressed the problem of temporal reasoning in the clinical domain by providing annotated clinical notes, pathology and radiology reports in line with Clinical TempEval challenges 2015/16, across two different evaluation phases focusing on cross domain adaptation. Our team focused on subtasks involving extractions of temporal spans and relations for which the developed systems showed average F-score of 0.45 and 0.47 across the two phases of evaluations.

1 Introduction

Temporal information extraction has been a widely explored topic of research interest in the field of information extraction during recent years. It is essential for improving the performance of applications such as question answering, search, text classification and systems that establish timelines from clinical narratives. In this line 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 2015/16 Clinical TempEval challenge (Bethard et al., 2015, 2016). These challenges provide annotated corpora on temporal entities and relations, which facilitate comparisons of multiple systems and push the state of art in the development of clinical temporal information extraction methodologies.

The 2017 Clinical TempEval challenge is the most recent community challenge that addresses temporal information extraction from clinical notes. The challenge was in inline with 2015/16 challenge in terms of subtasks. However this year's challenge focussed on cross domain adaptation across two phases of evaluation. In phase one (unsupervised domain adaptation), the systems were evaluated on their results for all six sub-tasks on brain cancer notes given colon cancer notes (data of 2015/16 challenge) as inputs. In phase two (supervised domain adaptation), evaluation was carried out in line with phase one but a small number of annotations of brain cancer notes were also given as inputs.

In this paper, we describe an end-to-end system that addresses subtasks involving extractions of temporal spans and relations. We designed the system by adapting various insights and techniques from previous work on temporal information extraction in the clinical domain (Sarath et al., 2016; Abdulsalam et al., 2016; Johri et al., 2014) and ensemble modelling (Dzeroski and Zenko, 2004).

The rest of the paper is organised as follows: In section 2, we discuss datasets, methods and feature sets used for each of the subtasks. In section 3, we present the results for various subtasks and conclude our work in section 4 with some of our findings and possible implications on future work.

2 Dataset and Methods

The THYME corpus (Styler et al., 2014) used in this task consists of clinical, pathology and radiology notes for colon/brain cancer patients from Mayo clinic (Bethard et al., 2017).

We designed an end-to-end pipeline consisting of four modules which process the input text in three stages: In stage one, the first and second

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 1005–1009,
Vancouver, Canada, August 3 - 4, 2017.

modules extract time/event expressions and their spans. In stage two, the third module predicts document time relations between the event and document creation time expressions. Finally, all the outputs of stage one and two are used to extract container relations in stage three. For phase one (unsupervised domain adaptation) we used train, dev and test colon cancer datasets to train and evaluated on the brain cancer test dataset. While in phase two we retrained models with a mixture of colon cancer and additional brain cancer notes, each of which are explained in upcoming sections. For our temporal information extraction system we used following open source libraries. 1) Stanford-CoreNLP (Manning et al., 2014) 2) scikit-learn (Pedregosa et al., 2011) 3) NLTK (Loper and Bird, 2002) 4) XGBoost (Chen and Guestrin, 2016) 5) Apache CTAKES (Savova et al., 2010) 6) ClearTK (Bethard et al., 2014) 7) H2O[1]

2.1 Time span identification

In the first stage our system identifies the time expressions and their spans.

Our manual observation of the colon cancer and brain cancer notes revealed that different time expressions show specific set of characteristics (Sarath et al., 2016) unique to each of the TIMEX3 class. Such characteristics may limit the systems 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 times as well as their types simultaneously.

Based on above observations and previous works on entity recognition tasks in the clinical domain (Lin et al., 2016), five Conditional Random Field(CRF) classifiers (Lafferty et al., 2001) were employed to identify each class of TIMEX3 expression except "duration" class for which we built a simple rule based system using Stanford TokensRegex Framework (Chang and Manning, 2014). For training CRF we tagged each token with either O (outside of a time mention), B-type (beginning of a time mention of type), or I-type (inside of a time mention of type), where type can be any of the TIMEX3 types defined by the Clinical TempEval challenge.

Features: n-grams (uni-, bi-) of nearby words (window size of +/- 2), character n-grams (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) and word shape features.

Unsupervised adaptation Run 1: CRF trained system only on colon cancer notes.

Supervised adaptation Run 1 & 2: Additionally we used additional 30 brain cancer notes.

2.2 Event span identification

Following the extraction of time expression, our system then identifies event expressions and their spans. Similar to time expression event expression also exhibited characterstic behaviour (Sarath et al., 2016; Abdulsalam et al., 2016).

A single CRF classifier was trained for extraction of event terms from clinical notes for which we used features that are described in section 2.1. Additionally we used following set of features.

Additional features: Word shape features of higher order, features showing disjunctions of words anywhere in the left or right, Conjoin of word shape and n-gram features. All the above features are described in StanfordCoreNLP (2014)

Both supervised and unsupervised adaptations differ as described previously in section 2.1.

2.3 Document time relation identification

Given spans of event mentions, our system further identifies relations between events and the document creation time using an NER (Named entity recognition) classifier trained for BEFORE, OVERLAP, BEFORE-OVERLAP or AFTER types.

Unsupervised adaptation run 1: Uses ClearTK NER chunking classifier (CRF) and features specified in section 2.1 extracted from the window of ± 5 words.

Supervised adaptation run 1: Similar to unsupervised adaptation run 1 except usage of additional 30 brain cancer notes.

Supervised adaptation run 2: Similar to supervised adaptation run 1 except ClearTK NER was replaced by a two-layer perceptron NER using H2O toolkit and skip-gram based word2vec word embeddings.

[1](http://www.h2o.ai/)

2.4 TLINK:Contains identification

We divide the task of narrative container relation identification into four 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 or within two adjacent sentences. For each sub-problem, we trained an different set of classifiers that identifies whether an ordered pair of two events/times (or a candidate pair) forms a TLINK of Contains type, using the scikit-learn package. Before training the classifiers, we apply the following steps in order to take into account the data distribution characteristics.

Firstly, 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 unlikely to form a TLINK:Contains relation based on the THYME corpus annotation guidelines (Lee et al., 2016) and heuristic rules (Abdulsalam et al., 2016). Secondly we apply cost sensitive learning in order to balance the effect of the larger negative samples present in the final set used for training. For each class we assigned weight proportional to class frequency.

Unlike event and time expressions, where we used single classifier such as CRF or a single multi layer perceptron network, for container relations, based on our previous experiences in relation extraction we used stacking of multiple classifiers to further reduce the effect of negative class overfitting. As such we used ensemble of Gradient boosted trees, XGBoost, Extra Trees (Geurts et al., 2006), Random forest (Breiman, 2001) classifier for extraction of narrative containers with following features. During model development all the hyperparamters were tuned using grid search with colon cancer notes (training) and 50% of brain cancer notes (validation).

Common features: Event/time tokens and its POS features, Special punctuation characters between event/time mentions, other event/time mentions within the same sentence, number of other event/time mentions between the two event/time mentions, verb tenses, section headers, sentence length.

Special features: A flag to indicate the presence of a pair in colon cancer data and a flag to indicate if the pairs were identified by pretrained CTAKES model.

Unsupervised adaptation Run 1: Stacked ensemble of gradient boosted decision trees, random forest, extra trees classifier with special features.

Supervised adaptation Run 1 & 2: Stacked ensemble of bagged XGBoost classifier, random forest and extra trees classifier re-trained on additional 30 brain cancer notes with event/time tokens and special features removed.

3 Experiments and Results

In this section, we present our system performance of various runs across two different phases for each of the participated subtasks. Tables 1-2 show the results of temporal span extraction and tables 3-4 shows results of temporal relation extraction. Our systems showed average F-score of 0.45 for unsupervised runs and 0.47 for supervised runs across different sub-tasks.

Submission runs	P	R	F
Unsupervised run 1	0.63	0.33	0.43
Supervised run 1 & 2	0.53	0.48	0.51

Table 1: Results of time expression

Submission runs	P	R	F
Unsupervised run 1	0.67	0.69	0.68
Supervised run 1 &2	0.67	0.75	0.71

Table 2: Results of event expression

Submission runs	P	R	F
Unsupervised run 1	0.44	0.45	0.45
Supervised run 1	0.49	0.55	0.52
Supervised run 2	0.42	0.47	0.44

Table 3: Results of doctime relation expression

3.1 Discussion

In this paper we described the system developed for temporal information extraction from clinical notes, using which we achieved average result of 0.45 for unsupervised and 0.47 for supervised phases of evaluation. We adapted state of the art techniques for entity recognition and relation extraction. We also experimented and evaluated stacked ensemble models involving XGBoost, Extra trees, Random Forest, Gradient Boosted trees for narrative container relation extraction.

Submission runs	P	R	F
Unsupervised run 1	0.23	0.22	0.23
Supervised run 1 & 2	0.11	0.27	0.15

Table 4: Results of narrative container relations

For time and event expression extraction our result (table 1 and table 2) were consistent across two phases and was on average 6% behind the best performing system. Potential reasons for the difference in F-score are i) Difference of our results with respect to gold standard annotation due to inclusion/exclusion of prepositions in certain expressions. For example, while a DURATION type time is annotated for the phrase "for the last 40 years" in the gold set, our system predicted a DURATION for the phrase "the last 40 years" omitting the preposition "for" from the gold standard annotation; ii) Limitations of features selected; iii) False negatives in event expression concerned with mispredictions in pathology and radiology reports; iv) Structural difference between colon and brain cancer notes, which is in agreement with improvement of results in phase two with the introduction of 30 brain cancer notes; In our future work we plan to investigate rule based methods to reduce preposition errors and filtering false negative. Further we plan to address problem of lesser training data of target domain through data augmentation techniques using deep learning methods.

For the document time relation extraction subtask, the CRF-based classification approach again allowed for significant improvements, particularly in phase two. Table 3 shows the evaluation scores obtained on the test set for DocTimeRel relation using CRF model. The final scores achieved in phase two (0.52) are comparable to the scores achieved (0.44) in phase one. This allows us to make consistent conclusions about classifier performance with and without supervision. Further when compared we could see that the top performing system had 5% higher F-score for DR task. A possible explanation for this and our future areas of concentration for improvement would be usage of different features related to the section where the event occurs, temporal expressions surrounding the event, and tense and aspect features of the predicates in the event context.

Narrative container relation extraction was the most difficult among all the subtasks as it suffers from major problem of data imbalance. For this work we employed pair/class weight selection strategies previously described in section 2.4 based on extensive experiments on colon cancer test set. Even though we tuned our system to achieve the results of the top performing system of clinical TempEval 2016 our system achieved very low result as shown in table 4. Our results are average of 13% behind the best performing system across two phases in CR task. Following are the major reason for this behaviour i) During testing the number of event-event pairs generated were very high, which made us to remove event-event pairs and submit only time-event pairs; ii) Removal of special features; iii) During phase two, bagging XGBoost resulted in overfitting of the model; iv) Also candidate pairs spanning across multiple sentences were missed by our classifier. During our experiments we observed most of the false positives followed pattern where both the expressions in pairs fall under same concept type in UMLS. Further we found that some pairs failed to satisfy parent child relationship in UMLS concept tree. Thus we plan to investigate rule based methods using UMLS that can identify and remove these kind of false positives. In addition to reducing false positives, this would also counteract against model overfitting when combined with a machine learning method. Also we believe further exploration of future engineering is needed to capture the pairs that span across multiple sentences.

4 Conclusion

Temporal information extraction from clinical notes remains a challenging task. Our analysis of different machine learning approaches have been informative, and resulted in competitive results for the 2017 Clinical TempEval subtasks. From our experiments we observe that CRF's generalize fairly well for extraction of time and event expressions. At the same time we can see there is a large room for improvement (methods and standardizations) in area of narrative container relations extraction. In future we plan to further improve our system to show higher performance based on the above observations.

Acknowledgments

We thank Mayo clinic and Clinical TempEval organizers for providing access to THYME corpus and other helps provided for our participation in the competition.

References

Abdulrahman Al Abdulsalam, Sumithra Velupillai, and Stéphane Meystre. 2016. Utahbmi at semeval-2016 task 12: Extracting temporal information from clinical text. In *SemEval@NAACL-HLT*.

Steven Bethard, Leon Derczynski, Guergana K. Savova, James Pustejovsky, and Marc Verhagen. 2015. Semeval-2015 task 6: Clinical tempeval. In *SemEval@NAACL-HLT*.

Steven Bethard, Philip V. Ogren, and Lee Becker. 2014. Cleartk 2.0: Design patterns for machine learning in uima. In *LREC*.

Steven Bethard, Guergana K. Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. Semeval-2016 task 12: Clinical tempeval. In *SemEval@NAACL-HLT*.

Steven Bethard, Guergana K. Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2017. Semeval-2017 task 12: Clinical tempeval. In *SemEval@ACL*.

Leo Breiman. 2001. Random forests. *Machine Learning* 45:5–32.

Angel X. Chang and Christopher D. Manning. 2014. Tokensregex: Defining cascaded regular expressions over tokens.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *KDD*.

Saso Dzeroski and Bernard Zenko. 2004. Is combining classifiers with stacking better than selecting the best one? *Machine Learning* 54:255–273.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine Learning* 63:3–42.

Nishikant Johri, Yoshiki Niwa, and Veera Raghavendra Chikka. 2014. Optimizing apache ctakes for disease/disorder template filling: Team hitachi in the share/clef 2014 ehealth evaluation lab. In *CLEF*.

John D. Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*.

Hee-Jin Lee, Hua Xu, Jingqi Wang, Yaoyun Zhang, Sungrim Moon, Jun Xu, and Yonghui Wu. 2016. Uthealth at semeval-2016 task 12: an end-to-end system for temporal information extraction from clinical notes. In *SemEval@NAACL-HLT*.

Chen Lin, Dmitriy Dligach, Timothy A. Miller, Steven Bethard, and Guergana K. Savova. 2016. Multilayered temporal modeling for the clinical domain. *JAMIA* 23:387–395.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *CoRR* cs.CL/0205028.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. pages 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010.

Danielle L. Mowery, Sumithra Velupillai, Brett R. South, Lee M. Christensen, David Martínez, Liadh Kelly, Lorraine Goeuriot, Noémie Elhadad, Sameer Pradhan, Guergana K. Savova, and Wendy W. Chapman. 2014. Task 2: Share/clef ehealth evaluation lab 2014. In *CLEF*.

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.

Sarath, Manikandan R, and Yoshiki Niwa. 2016. Hitachi at semeval-2016 task 12: A hybrid approach for temporal information extraction from clinical notes. In *SemEval@NAACL-HLT*.

Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin Kipper Schuler, and Christopher G. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *JAMIA* 17:507–513.

Scharolta Katharina Siencnik. 2015. Adapting word2vec to named entity recognition. In *NODALIDA*.

Stanford-CoreNLP. 2014. Stanford ner feature factory. https://tinyurl.com/zanvz7c.

William F. Styler, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C. de Groen, Bradley James Erickson, Timothy A. Miller, Chen Lin, Guergana K. Savova, and James Pustejovsky. 2014. Temporal annotation in the clinical domain. *TACL* 2:143–154.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *JAMIA* 20.