GUIR at SemEval-2017 Task 12: A Framework for Cross-Domain Clinical Temporal Information Extraction

Sean MacAvaney, Arman Cohan, and Nazli Goharian
Information Retrieval Lab
Department of Computer Science
Georgetown University
{firstname}@ir.cs.georgetown.edu

Abstract

Clinical TempEval 2017 (SemEval 2017 Task 12) addresses the task of cross-domain temporal extraction from clinical text. We present a system for this task that uses supervised learning for the extraction of temporal expression and event spans with corresponding attributes and narrative container relations. Approaches include conditional random fields and decision tree ensembles, using lexical, syntactic, semantic, distributional, and rule-based features. Our system received best or second best scores in TIMEX3 span, EVENT span, and CONTAINS relation extraction.

1 Introduction

Clinical TempEval 2017 (Bethard et al., 2017) is designed to address the challenge of extracting clinical timelines from medical narratives. It is a successor to Clinical TempEval 2016 (Bethard et al., 2016), Clinical TempEval 2015 (Bethard et al., 2015), and the i2b2 temporal challenge (Sun et al., 2013).

Clinical TempEval evaluates systems using the THYME corpus (Styler IV et al., 2014), which is annotated with temporal expressions (TIMEX3), events (EVENT), and temporal relations (TLINK) per an extension of the TimeML specifications (Pustejovsky et al., 2003).

The focus of Clinical TempEval 2017 is domain adaptation. The source domain consists of clinical text about patients undergoing colon cancer treatments, while the target domain consists of clinical text about those with brain cancer. There are two phases in the task. In phase 1, the shared task provides no annotations for the target domain (unsupervised). In phase 2, the shared task provides a small annotated training set from the target domain (supervised). Both phases evaluate system performance on thirteen tasks via precision, recall, and F1-score.

In Clinical TempEval 2016, the top-performing system employed structural support vector machines (SVM) for entity span extraction and linear support vector machines for attribute and relation extraction (Lee et al., 2016). For the previous iteration, Velupillai et al. (2015) developed a pipeline based on ClearTK and SVM with lexical and rule-based features to extract TIMEX3 and EVENT mentions. In the i2b2 2012 temporal challenge, all top performing teams used a combination of supervised classification and rule-based methods for extracting temporal information and relations (Sun et al., 2013). Other efforts in clinical temporal annotation include works by Roberts et al. (2008), Savova et al. (2009), and Galescu and Blaylock (2012).

Previous work has also investigated extracting temporal relations. Examples of these efforts in the general domain include: classification by SVM (Chambers et al., 2007), Integer Linear Programming (ILP) for temporal ordering (Chambers and Jurafsky, 2008), Markov Logic Networks (Yoshikawa et al., 2009), and SVM with Tree Kernels (Miller et al., 2013).

In this paper, we present a framework for temporal information extraction in clinical narratives. Specifically we utilize Conditional Random Fields (CRFs) and decision tree ensembles for extracting temporal entities and relations from clinical text. The features we use are covered in detail in Section 2. This work can be seen as an extension and refinement of the system used for Clinical TempEval 2016 by Cohan et al. (2016).

2 Methodology

Our approach uses supervised learning algorithms with lexical, syntactic, semantic, distributional, and rule-based features for span, attribute, and relation extraction.

2.1 Span Extraction

Extraction of TIMEX3 and EVENT spans uses linear-chain CRFs.

We use BIO labels for the classification of spans of text from the tokenized source text: ”B” indicates that the token begins a span, ”I” indicates that the token is inside the span, and ”O” indicates that the token is outside all spans. This approach allows for spans to represent one or more adjacent tokens. Non-contiguous spans, although not supported, have a low occurrence.

Basic lexical features computed for each token are as follows: lowercase form of the token; uppercase and lowercase flags; prefix and suffix; lemmatized form; shape; punctuation flag; and stop word flag. Syntactic features are coarse- and fine-grained part-of-speech tags. We used spaCy[1] for tokenization and basic features. In addition, we used the Unified Medical Language System (UMLS) ontology (Bodenreider, 2004) via MetaMap[2] to capture semantic concepts and use them as features. We limited the types to those indicative of clinical events (diagnostic procedure, disease or syndrome, and therapeutic procedure).

We also include regular expression-based features to capture more complicated and specialized token properties (summarized in Table 1). While the more generalized features we used (e.g. shape and suffix) capture some of the same information, this approach prioritizes likely generalizations and avoid over-fitting to specific cases. For instance, it allows the algorithm to generalize “Summer 2010” as “[Season] [Year]” instead of a more literal sequence.

We use distributional features for generalization. We construct Brown clusters (Brown et al., 1992) on the text with fifty clusters. The binary representation of each token’s cluster is a feature. We also use word embeddings trained using Word2Vec (Mikolov et al., 2013) on the MIMIC-III dataset (Johnson et al., 2016) with a dimension of 100. The word embeddings also encode token usage context, and thus should generalize the model.

Feature	Examples
Date	12/3/2010, 1965-01-21
Month	January, Aug
Day	1st, 31
Day-of-week	Monday, Wed
Season	summer, spring
Year	2013, 1990s
Time	8:42, a.m.
Time Unit	minute, sec
Number	4, seventeen
Temporal preposition	in, after
Temporal adverb	daily, lately
Temporal prefix	pre, post

Table 1: Rule-based features and examples.

For each token’s feature set, we also include the features from the ±1 adjacent tokens.

2.2 Attribute Extraction

We treat the extraction of the attributes of EVENT and TIMEX3 as a classification problem. Our system trains a CRF model for each attribute, with the labels of each model corresponding to the attribute values and the same features used in span extraction. An expanded window of ±3 tokens is used for this task. Our system treats DOCTIMEREL (the EVENT’s temporal relation to the document time) as attribute extraction.

2.3 Narrative Containers

Our approach trains gradient boosted trees (Friedman, 2001) on candidate relation pairs and uses this model to predict relations. Our system uses XGBoost (Chen and Guestrin, 2016) for this task.

Clinical TempEval 2017 only considers temporal links (TLINK) with a type of CONTAINS; other types of TLINKs are not evaluated due to lower inter-annotator agreements. Our system uses TLINK type labels when the relation exists, and a null label when the candidate relation does not represent an actual relation. We note that our approach extracts all relation types. Our system uses both entity features (describing each relation endpoint) and relation features (describing the relationship between the source and target).

Entity features include the entity type, entity attribute values, and the case-folded text value. Additionally, we use each token and related features (e.g. suffix) contained within the entity as features. We apply semantic Role Labeling (SRL) to the sentence containing the entity, which identifies semantic predicates in the sentence per PropBank guidelines (Palmer et al., 2005). If the entity text

[1] spacy.io

[2] https://metamap.nlm.nih.gov/; 2016 version

	TIMEX3 Spans			EVENT Spans		
	P	R	F1	P	R	F1
Phase 1 (Unsupervised)						
Our System	0.61	0.53	†**0.57**	0.64	‡**0.80**	‡**0.71**
Median	0.63	0.46	0.48	0.64	0.69	0.68
Phase 2 (Supervised)						
Our System	‡**0.57**	‡**0.62**	†**0.59**	‡**0.68**	0.82	‡**0.74**
Median	0.53	0.52	0.54	0.67	0.76	0.71
MEMORIZE	0.64	0.22	0.33	0.61	0.51	0.56

Table 2: Evaluation results for span extraction. †Top score. ‡Second highest score.

is found in a semantic predicate, we use the argument label as a feature for the model. We used the SENNA[3] implementation for SRL tagging.

Relation features capture information about the relationship between two entities. Basic relation features included are the character distance between the entities and the pair of entity types. Syntactic features applied capture the path along the constituent and dependency trees between the entities. Our system uses the spaCy toolkit for dependency parsing. We derive n-gram segments of the path, the full path, and the distance of the path, and use them as features.

We limit candidate relations to permutations of entities belonging to the same sentence. This approach precludes relations that cross sentence boundaries, but limits the extent of negative training samples.

2.4 Domain Adaptation

Our system splits the phase 2 text ("train10") into a dev set and a test set. A grid search is performed for span, property, and relation extraction over the applicable hyperparameters. Text from the source domain is used for training, and the dev set from the target domain is used for evaluation. The test set is used after the grid search to verify that the procedure did not overfit hyperparameters.

3 Experimental Setup

In phase 1, we train our system on all available annotations from the source domain. In phase 2, we train our system on all available data from the source domain and the "train10" dataset from the target domain.

Baselines The baselines are two rule-based systems (Bethard et al., 2015) that the shared task provides along with the corpus. The MEMORIZE baseline, which is the baseline for all tasks except for narrative containers, memorizes the EVENT and TIMEX3 mentions and attributes based on the training data. Then it uses the memorized model to extract temporal information from new data. For narrative containers, the CLOSEST baseline predicts a TLINK relation with type CONTAINS between every TIMEX3 annotation and its closest EVENT.

Furthermore, we compare our results against the other submissions to Clinical TempEval 2017. We report the median value for each metric, as well as indicators when our system achieves either the top result (†), or second-highest result (‡). Only the systems that submitted values for a particular task are considered; systems reported as $p = 0.00$, $r = 1.00$, and $F1 = 0.00$ are ignored.

Evaluation metrics Clinical TempEval 2017 evaluates thirteen tasks. Each task reports the precision recall, and F1-score of the submitted results as compared to a human annotated and adjudicated ground truth. The following tasks are not reported in this paper for brevity: "All spans & all properties", "All spans only", "Time span & all properties", and "Event span & all properties".

4 Results and discussion

Our system outperformed other participating systems, receiving best or second best results extracting TIMEX3 spans, EVENT spans, and CONTAINS relations. Generally our domain adaptation procedure improved results, but it reduced the results of CONTAINS relations. Although we received top scores, we fell short of the single-domain performance achieved in Clinical TempEval 2016.

Table 2 shows the results for TIMEX3 and EVENT span extraction. Our system achieved the top F1 score for TIMEX3 spans and the second highest F1 score for EVENT spans in both phases. Furthermore, our system met or exceeded the median and MEMORIZE baseline in all but one metric (TIMEX3 precision), in which it had significant gains in recall. Table 3 shows the results for TIMEX3 and EVENT attribute extraction. We note that while our system performs well on some of these categories, on some other categories it underperforms the median results (e.g. such as EVENT Modality and EVENT Polarity). Our system performed well at CONTAINS relations, but only achieved median results at DOCTIMEREL

[3] http://ml.nec-labs.com/senna/

	TIMEX3 Class			EVENT Modality			EVENT Degree			EVENT Polarity			EVENT Type		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Phase 1 (Unsupervised)															
Our System	0.55	0.47	‡**0.51**	0.50	0.64	0.56	0.61	‡**0.77**	‡**0.68**	0.59	‡**0.74**	0.65	0.61	‡**0.76**	‡**0.68**
Median	0.56	0.45	0.46	0.55	0.63	0.59	0.62	0.71	0.68	0.60	0.70	0.66	0.61	0.70	0.66
Phase 2 (Supervised)															
Our System	‡**0.54**	‡**0.59**	†**0.56**	0.60	0.72	‡**0.66**	‡**0.67**	0.80	‡**0.73**	0.54	0.64	0.58	‡**0.66**	0.79	‡**0.72**
Median	0.49	0.48	0.48	0.57	0.68	0.63	0.66	0.77	0.71	0.62	0.70	0.66	0.65	0.76	0.70
MEMORIZE	0.49	0.17	0.25	0.29	0.24	0.26	0.47	0.40	0.43	0.56	0.47	0.51	0.50	0.42	0.46

Table 3: Evaluation results for attribute extraction. †Top score. ‡Second highest score.

	CONTAINS			DOCTIMEREL		
	P	R	F1	P	R	F1
Phase 1 (Unsupervised)						
Our System	†**0.52**	0.25	†**0.34**	0.36	0.45	0.40
Median	0.33	0.25	0.32	0.39	0.45	0.41
Phase 2 (Supervised)						
Our System	†**0.59**	0.16	0.25	0.45	0.55	0.50
Median	0.20	0.16	0.16	0.42	0.51	0.46
CLOSEST	0.33	0.08	0.12	-	-	-
MEMORIZE	-	-	-	0.22	0.18	0.20

Table 4: Evaluation results for relation extraction. †Top score.

relations (see Table 4). In phase 1, our system achieved the top results for CONTAINS precision and F1. Our domain adaptation procedure resulted in a drop in recall for CONTAINS relations. We suspect this is due to overfitting the model to the sample data. We suspect that including more contextual or semantic features would improve the performance of attribute extraction (including DOCTIMEREL).

4.1 Error Analysis

We conducted an unsupervised domain adaptation run against the “train10” dataset to get an idea of failure cases. (We could not use the full target domain test set because these data are not available.)

One issue with TIMEX3 extraction is previously unseen or atypical date formats, for instance “12Jun2013” (no hyphens). One way to resolve this issue could be to use a more generalized library for extracting time expressions (e.g. HeidelTime), but even this library does not extract the example shown above. Furthermore, it would not generalize to new and otherwise unknown formats. The supervised training subset could be used in each domain to identify these kinds of conventions, but this is labor-intensive and prone to error. Another issue is inconsistency in TIMEX3 annotation conventions (e.g. annotating a date and time separately sometimes and jointly in others). This complicates the model and leads to otherwise inexplicable annotation absences.

One example of an EVENT extraction failure is the false positive of “Cancer” in the phrase “Cancer Research Hospital”. An approach to resolve this would be to use named entity recognition features, or by treating named entities as chunks that are annotated using a different technique. False positive EVENTs were common in certain sections of the notes (e.g. ongoing care; suggested interventions), indicating that document segmentation by section could be useful. This would only work in a supervised environment, unless domain sections have a great degree of overlap and can be mapped to one another.

TLINK error cases include the known limitation of intra-sentence relations. Other false negatives candidates seemed to be due to domain-specific language (e.g. “temozolomide”), suggesting that lexical features are overused, or the syntactic and semantic features we use are inadequate.

5 Conclusions

The results of Clinical TempEval 2017 show that there is still room to explore cross-domain temporal information extraction. We presented a system for both unsupervised and supervised temporal domain adaptation. It performed among best of participating teams, receiving best or second best scores in TIMEX3 span, EVENT span, and CONTAINS relation extraction. All teams fell short of meeting the top results for the source domain. Future work in this area could focus on techniques for using a small number of annotations to tune a system to other domains due to the modest improvements in phase 2.

References

Steven Bethard, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. Semeval-2015 task 6: Clinical tempeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics Denver, Colorado, pages 806–814.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. Semeval-2016 task 12: Clinical tempeval. *Proceedings of SemEval* pages 1052–1062.

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. Semeval-2017 task 12: Clinical tempeval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 563–570. http://www.aclweb.org/anthology/S17-2093.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research* 32(suppl 1):D267–D270.

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.

Nathanael Chambers and Dan Jurafsky. 2008. Jointly combining implicit constraints improves temporal ordering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 698–706.

Nathanael Chambers, Shan Wang, and Dan Jurafsky. 2007. Classifying temporal relations between events. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, pages 173–176.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pages 785–794.

Arman Cohan, Kevin Meurer, and Nazli Goharian. 2016. Guir at semeval-2016 task 12: Temporal information processing for clinical narratives. *Proceedings of SemEval* pages 1248–1255.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* pages 1189–1232.

Lucian Galescu and Nate Blaylock. 2012. A corpus of clinical narratives annotated with temporal information. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. ACM, pages 715–720.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Liwei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data* 3.

Hee-Jin Lee, Yaoyun Zhang, Jun Xu, Sungrim Moon, Jingqi Wang, Yonghui Wu, and Hua Xu. 2016. Uthealth at semeval-2016 task 12: an end-to-end system for temporal information extraction from clinical notes. *Proceedings of SemEval* pages 1292–1297.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Timothy Miller, Steven Bethard, Dmitriy Dligach, Sameer Pradhan, Chen Lin, and Guergana Savova. 2013. Discovering temporal narrative containers in clinical text. In *Proceedings of the 2013 Workshop on Biomedical Natural Langua ge Processing*. pages 18–26.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics* 31(1):71–106.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*. volume 2003, page 40.

Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Andrea Setzer, and Ian Roberts. 2008. Semantic annotation of clinical text: The clef corpus. In *Proceedings of the LREC 2008 workshop on building and evaluating resources for biomedical text mining*. pages 19–26.

Guergana Savova, Steven Bethard, F William IV, IV Styler, James H Martin, Martha Palmer, James J Masanz, and Wayne H Ward. 2009. Towards temporal relation discovery from the clinical narrative. In *AMIA*. pages 568–572.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics* 2:143–154.

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.

Sumithra Velupillai, Danielle L Mowery, Samir Abdelrahman, Lee Christensen, and Wendy W Chapman. 2015. Blulab: Temporal information extraction for the 2015 clinical tempeval challenge. In *The 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado, June 4-5, 2015*. Association for Computational Linguistics, pages 815–819.

Katsumasa Yoshikawa, Sebastian Riedel, Masayuki Asahara, and Yuji Matsumoto. 2009. Jointly identifying temporal relations with markov logic. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*. Association for Computational Linguistics, pages 405–413.