Improving Temporal Relation Extraction with Training Instance Augmentation

Chen Lin and **Timothy Miller**
Boston Children's Hospital Informatics Program
and Harvard Medical School
{first.last}@childrens.harvard.edu

Dmitriy Dligach
Loyola University Chicago
ddligach@luc.edu

Steven Bethard
University of Alabama at Birmingham
bethard@uab.edu

Guergana Savova
Boston Children's Hospital Informatics Program
and Harvard Medical School
guergana.savova@childrens.harvard.edu

Abstract

Temporal relation extraction is important for understanding the ordering of events in narrative text. We describe a method for increasing the number of high-quality training instances available to a temporal relation extraction task, with an adaptation to different annotation styles in the clinical domain by taking advantage of the Unified Medical Language System (UMLS). This method notably improves clinical temporal relation extraction, works beyond featurizing or duplicating the same information, can generalize between-argument signals in a more effective and robust fashion. We also report a new state-of-the-art result, which is a two point improvement over the best Clinical TempEval 2016 system.

1 Introduction

Temporal relation extraction is important for understanding ordering of events from a narrative text. Recent years have seen annotated corpora created for temporal information extraction, from newspaper text (Pustejovsky et al., 2003; Verhagen et al., 2007; Verhagen et al., 2010; UzZaman et al., 2013), to clinical narratives (Savova et al., 2009; Sun et al., 2013; Styler et al., 2014), all with the aim of developing systems for building event timelines from textual descriptions of events. Such narrative timelines are important for information extraction tasks such as question answering (Kahn et al., 1990), clinical outcomes prediction (Schmidt et al., 2005; Lin et al., 2014), and the identification of temporal patterns (Zhou and Hripcsak, 2007) among many.

In a typical supervised approach to the temporal relation extraction task, argument pairs consist of pairs of events or temporal expressions. Corpora differ in their syntactic annotation of such expressions. For example, the THYME corpus, consisting of oncology, pathology and radiology notes, annotated only event headwords (Styler et al., 2014), while the i2b2 corpus, consisting of discharge summaries, annotated entire noun phrases as events (Sun et al., 2013). As a result, it is necessary to account for these differences when implementing a generalizable relation extraction system.

However, the annotations of the temporal relations between the events remain unaffected by the choice of headwords or phrases for the event annotation. For example, in a relation between the temporal expression *yesterday* and the event *severe lower abdominal pain*, if the argument had been the head word *pain* it still would have been an instance of the same temporal relation. Thus, we can automatically create additional training examples by varying the extent of headword expansion. For example, the relation between *yesterday* and *severe lower abdominal pain* can automatically generate four valid relations of the same type where the second arguments are *pain*, *abdominal pain*, and *lower abdominal pain*.

In this paper, we describe an automatic method that generates more temporal training instances by semantically expanding gold medical events based on a clinical ontology, the Unified Medical Language System (UMLS) (Lindberg et al., 1993). It bridges the gap between different syntactic annotations of events in clinical corpora. We show that this method is superior to representing the same information as additional features, that it differs from plain upsampling, and that the primary mechanism of improvement is in the better representation of between-argument features. Our method can be viewed as a new form of data augmentation, akin to the generation of image variants for vision recognition (Krizhevsky et al., 2012) or the generation of word substitutions for information extraction (Kolomiyets et al., 2011).

Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pages 108–113,
Berlin, Germany, August 12, 2016. ©2016 Association for Computational Linguistics

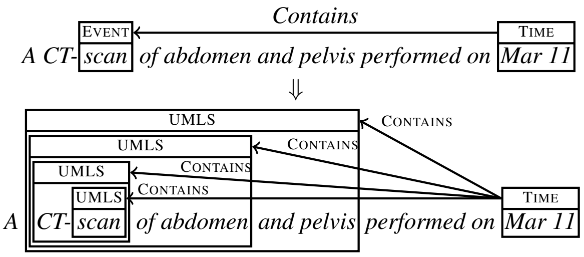
Contains
EVENT
TIME
A CT-scan of abdomen and pelvis performed on Mar 11
⇓
UMLS
CONTAINS
UMLS
CONTAINS
UMLS
CONTAINS
UMLS
CONTAINS
TIME
A CT-scan of abdomen and pelvis performed on Mar 11

Figure 1: Example expansion of the event "scan"

2 Method

First, the text was scanned for any medical concepts from the UMLS Metathesaurus (`http://www.nlm.nih.gov/research/umls/`), a collection of concepts from different biomedical terminologies. Apache cTAKES (`http://ctakes.apache.org`) was used to extract such UMLS concepts. Next, we use these UMLS concepts and gold standard events to expand relation arguments. For a gold standard event e annotated by the headword, we define $\textsc{Expand}(e)$ as the set of UMLS entities whose spans cover e. If e is involved in a temporal relation r, we assume u ($u \in \textsc{Expand}(e)$) is involved in the same relation and therefore we generate a new temporal relation that is identical to r but with the event e replaced by a UMLS entity u. Figure 1 shows an example of expanding the gold event "scan" to its covering UMLS entities and generating related relations.

We differentiate temporal relations into event-time and event-event, and expand relations as detailed in Algorithm 1 and Algorithm 2, respectively. For event-event, we ensure the event spans do not overlap after expansion. Our event-time model classifies all relations – CONTAINS, BEFORE, OVERLAP, BEGINS-ON, ENDS-ON and NONE, while our event-event model classifies only CONTAINS and NONE relations due to the very low inter-annotator agreement for the other relation types in our evaluation corpus (Styler et al., 2014). For both models, NONE is used to indicate that there is no relation between a pair of arguments.

Algorithm 1 Expansion for event-time relations

```
Given a gold-standard annotated event-time relation $r(e,t)$, where $e$ is an event, $t$ is a temporal expression, $r \in \{\textsc{Contains}, \textsc{Before}, \ldots, \textsc{None}\}$
for UMLS entity $u \in \textsc{Expand}(e)$ do
    Create relation $r'(u, t)$, $r' \leftarrow r$
    Add $r'$ to training data
end for
```

Algorithm 2 Expansion for event-event relations

```
Given a gold-standard annotated event-event relation $r(e_a,e_b)$, where $e_a$, $e_b$ are events, $r \in \{\textsc{Contains}, \textsc{None}\}$
for UMLS entity $u_a \in \textsc{Expand}(e_a)$ do
    if not overlaps(span($u_a$), span($e_b$)) then
        Create relation $r'(u_a, e_b)$, $r' \leftarrow r$
        Add $r'$ to training data
    end if
end for
for UMLS entity $u_b \in \textsc{Expand}(e_b)$ do
    if not overlaps(span($e_a$,$u_b$)) then
        Create relation $r'(e_a, u_b)$, $r' \leftarrow r$
        Add $r'$ to training data
    end if
end for
```

3 Experiments

3.1 Dataset

We tested our event expansion technique on a publicly available clinical corpus: the colon cancer set of the THYME corpus (Styler et al., 2014) used in SemEval 2015 Task 6 (Bethard et al., 2015) and SemEval 2016 Task 12: Clinical TempEval (Bethard et al., 2016). It contains 600 documents (400 oncology notes and 200 pathology notes) of 200 colon cancer patients. The gold standard annotations contain events (including both medical and general events, all annotated by head words), temporal expressions (e.g. *tomorrow*, *postoperative*, and *March-11-2009*), and temporal relations. We used the same training/development/test split as Clinical TempEval. The development set was used for testing research questions and building final models. Once the models were deemed finalized, they were rebuilt on the combined training and development sets and tested on the test set.

3.2 Models

We built two within-sentence temporal-relation classification models, one for event-time relations and one for event-event relations. We paired every gold event with every gold time expression within the same sentence to form candidate instances for the event-time classifier. We paired all gold events within a sentence to form candidates for the event-event classifier. For training, gold relations were also expanded by calculating the closure sets of all possible relations in a clinical document.

We use the LIBLINEAR (Fan et al., 2008) L2-regularized L2-loss dual SVM as the learning algorithm for both models. Our features for event-time and event-event models are shown in Table 1.

Feature	Description	EE	ET
Tokens	the first and the last word of each concept, all words covered by a concept as a bag, bag-of-words around each concept for a window of [-3, 3], bag-of-words between two concepts, and the number of words between two concepts	✓	✓
Part-of-speech tags	the POS tags of each concept as a bag	✓	
Event attributes	all event-related attributes such as polarity, modality, and type	✓	✓
UMLS feature	UMLS semantic types as features	✓	
Dependency path	the dependency path between two concepts and the number of dependency nodes in-between	✓	✓
Overlapped head	if two concepts share the same head word	✓	
Temporal attributes	the class type of a time expression, e.g. Date, Time, Duration, etc.		✓
Special words	Any words from the time lexicon developed by NRCC (Cherry et al., 2013a) that the concepts or the context in-between contain		✓
Nearest flag	if the event-time pair in question is the closest among all pairs in the same sentence		✓
Conjunction feature	if there is any conjunction word between the arguments		✓

Table 1: Features used for event-event (EE) and event-time (ET) classifiers

3.3 Research questions

We investigate the following questions:

1. Can the effect of the UMLS expansion technique be replicated using additional features? One may wonder if adding instances via UMLS expansion is isomorphic to adding more features that capture the UMLS information. To answer this question, we find all covering UMLS entities, but instead of creating new instances, extract token features from these entities and add those to the other features for the instance.

2. Is it better to expand to the longest UMLS entity or to expand to all possible spans? In our Figure 1 example, the longest UMLS entity covering "scan" is "CT-scan of abdomen and pelvis". But we could also create instances for the UMLS entities "scan", "CT-scan" and "CT-scan of abdomen". We also compare against a purely linguistic expansion to the immediate enclosing noun phrase (NP).

3. Is the improvement due to the replication of instances? Our expansion technique creates many similar relations, and in cases where a UMLS entity has the same span as a gold event, the technique creates true duplicate instances. For example, the relation CONTAINS(scan, March 11) is duplicated in Figure 1. Thus we also compare our UMLS-informed expansion of instances to simple duplication of instances[1].

4. Which types of features benefit most from the expansion? There are three groups of token features: within each argument, between the arguments, and the preceding and following three words (context) of an argument. We test the performance one feature group at a time, with and without the event expansion.

We test all research questions by training on the training set and testing on the development set with token-based features for the event-time relations. Note that expansion is applied only to the training set, not to the development or test set.

3.4 Evaluation

For results on the development set, we calculate closure-enhanced precision, recall and F1-score (UzZaman and Allen, 2011) on just the within-sentence relations (since that's what our models are able to predict). Precision is the percentage of system-generated relations that can be verified in the transitive closure of the gold standard relations. Recall is the percentage of gold standard relations that can be found in the transitive closure of the system-generated relations. The final F1-score is the harmonic mean of the transitive-closure-processed precision and recall.

For results on the test set, we used the official Clinical TempEval evaluation scripts so that our results are directly comparable with the outcomes of Clinical TempEval 2016 (Bethard et al., 2016). These scripts use similar definitions of closure-enhanced precision, recall and F1-score, but evaluate only CONTAINS relations in oncology notes.

4 Results on the development set

Question 1 is answered by the first two rows of Table 2: adding token features representing expanded UMLS entities does not achieve the same perfor-

[1] In SVM classification, duplicating training instances can affect the cost penalty by altering the number of instances within the margin. It is thus critical to tune cost parameter C for all experiments, which we do on development data.

P	R	F	#instances	Settings
0.587	0.538	0.561	8423	no Expansion
0.466	0.455	0.460	8423	UMLS as features
0.578	0.533	0.555	16846	duplicate instances
0.580	0.534	0.556	25269	triple instances
0.605	0.557	0.580	9506	longest UMLS
0.592	0.592	0.592	10705	expand to NPs
0.654	0.591	0.621	12966	all UMLS

Table 2: Results on the development set. No expansion vs. encoding UMLS as features; duplicating and triplicating training instances; expand to the longest UMLS span, expand to the immediate enclosing NP vs. expand to all UMLS spans.

P	R	F	#instances	Settings
0.359	0.155	0.217	8423	(A) no expansion
0.582	0.206	0.304	12966	(A) with expansion
0.087	0.116	0.099	8423	(B) no expansion
0.600	0.546	0.572	12966	(B) with expansion
0.587	0.254	0.355	8423	(C) no expansion
0.648	0.264	0.375	12966	(C) with expansion

Table 3: Results on the development set. Comparison of improvement for feature groups: (A) words covered by the arguments; (B) words in between the arguments; (C) words around the arguments.

mance as UMLS expansion, and in fact decreases performance. Question 2 is addressed in the last three rows: expanding to all possible UMLS spans works better than expanding only to the longest span or to the immediate enclosing NP. Expanding to NPs achieved the second best result, suggesting that when a domain-specific ontology is unavailable, expansion via syntax might provide a viable alternative. Question 3 is answered by rows 1, 3 and 4: when the cost parameter is properly tuned, doubling or tripling instances (rows 3 and 4) does not improve performance over no expansion (row 1). Question 4 is addressed by Table 3: features extracted between the two arguments achieve the biggest gain from our expansion method.

5 Results on the test set

Once the parameters were fine-tuned, we trained both event-time and event-event models on the combined training and developments sets, and tested them on the test set. All features described in Table 1 are used. The first two rows of Table 4 evaluate both event-event and event-time models, the next two rows evaluate only the event-time model, and the last two rows evaluate only the event-event model. Statistical significance is computed via Wilcoxon signed-rank tests over document-by-document comparisons, as in (Cherry et al., 2013b).

P	R	F	Settings	P-value
0.635	0.549	0.589	(1) no Expansion	0.117
0.669	0.534	0.594	(1) with Expansion	
0.673	0.291	0.407	(2) no Expansion	
0.708	0.287	0.408	(2) with Expansion	
0.594	0.252	0.354	(3) no Expansion	
0.628	0.243	0.351	(3) with Expansion	

Table 4: Results on the test set with all features. (1) Evaluate both Event-Time and Event-Event models; (2) Evaluate Event-Time model only; (3) Evaluate Event-Event model only. See Section 3.4 for explanation for why shaded scores are different from their counterparts in Table 2.

6 Discussion

Our experiments show our method is helpful for the event-time model, and not harmful for the event-event model. We hypothesize that the multiple instances capture the important surrounding context between arguments and allow more generalization over it. For the example in Figure 1, the most important features are "performed on." Our method weeds out less discriminative features by strengthening the important contextual signals that appear across many different entity boundaries. This is supported by the results of Table 3 (B). We suspect that the small improvement seen on the test data may be a result of the additional development examples canceling the benefit of augmented examples. This suggests that this method may be most effective in tasks with limited training instances.

Event-event relations are more complicated, first in that they have lower annotation quality than event-time relations (see Table 5 from (Styler et al., 2014)). And while almost every temporal expression in a sentence is important, not all events in a sentence are, creating many potential "distractor" events (e.g., *showed*) in the context of the clinical domain. We performed some exploratory experiments (not shown), restricting the data to only adjacent medical events in notes with high inter-annotator agreement, and saw significant performance improvements. But further study is needed to generalize this to all event-event relations.

With the presented method, our temporal relation system achieved F1 of 0.594, a two percentage-point improvement over the best Clinical TempEval 2016 system's F1 of 0.573 (Bethard et al., 2016). Our results also suggest that gains may be possible in the general domain by using syntactic constituents for expansion. The method is available open source at the temporal module of Apache cTAKES[2] (Savova et al., 2010).

7 Acknowledgements

Thanks to Sean Finan for technically supporting the experiments. The study was funded by R01 LM 10090 (THYME), R01GM103859 (PGx), and U24CA184407 (DeepPhe). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

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)*, pages 806–814, Denver, Colorado, June. Association for Computational Linguistics.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. Semeval-2016 task 12: Clinical tempeval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

Colin Cherry, Xiaodan Zhu, Joel Martin, and Berry de Bruijn. 2013a. À la recherche du temps perdu: extracting temporal relations from medical text in the 2012 i2b2 NLP challenge. *Journal of the American Medical Informatics Association*, 20(5):843–848.

Colin Cherry, Xiaodan Zhu, Joel Martin, and Berry de Bruijn. 2013b. la recherche du temps perdu: extracting temporal relations from medical text in the 2012 i2b2 nlp challenge. *Journal of the American Medical Informatics Association*, 20(5):843–848.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

Michael G Kahn, Larry M Fagan, and Samson Tu. 1990. Extensions to the time-oriented database model to support temporal reasoning in medical expert systems. *Methods of Information in Medicine*, 30(1):4–14.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2011. Model-portability experiments for textual temporal analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 271–276, Portland, Oregon, USA, June. Association for Computational Linguistics.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Chen Lin, Elizabeth W Karlson, Dmitriy Dligach, Monica P Ramirez, Timothy A Miller, Huan Mo, Natalie S Braggs, Andrew Cagan, Vivian Gainer, Joshua C Denny, and Guergana K Savova. 2014. Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. *Journal of the American Medical Informatics Association*.

Donald A Lindberg, Betsy L Humphreys, and Alexa T McCray. 1993. The unified medical language system. *Methods of information in Medicine*, 32(4):281–291.

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.

Guergana Savova, Steven Bethard, Will Styler, James Martin, Martha Palmer, James Masanz, and Wayne Ward. 2009. Towards temporal relation discovery from the clinical narrative. In *AMIA annual symposium proceedings*, volume 2009, page 568. American Medical Informatics Association.

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. *J Am Med Inform Assoc*, 17(5):507–513.

Reinhold Schmidt, Stefan Ropele, Christian Enzinger, Katja Petrovic, Stephen Smith, Helena Schmidt, Paul M Matthews, and Franz Fazekas. 2005. White matter lesion progression, brain atrophy, and cognitive decline: the austrian stroke prevention study. *Annals of neurology*, 58(4):610–616.

William Styler, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet 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*, 2:143–154.

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.

Naushad UzZaman and James F Allen. 2011. Temporal evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 351–356. Association for Computational Linguistics.

[2] `http://ctakes.apache.org`

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 75–80. Association for Computational Linguistics.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62. Association for Computational Linguistics.

Li Zhou and George Hripcsak. 2007. Temporal reasoning with medical dataa review with emphasis on medical natural language processing. *Journal of biomedical informatics*, 40(2):183–202.