GUIR at SemEval-2016 task 12: Temporal Information Processing for Clinical Narratives

Arman Cohan, Kevin Meurer and **Nazli Goharian**
Information Retrieval Lab
Department of Computer Science
Georgetown University
{arman, nazli}@ir.cs.georgetown.edu, kam346@georgetown.edu

Abstract

Extraction and interpretation of temporal information from clinical text is essential for clinical practitioners and researchers. SemEval 2016 Task 12 (Clinical TempEval) addressed this challenge using the THYME[1] corpus, a corpus of clinical narratives annotated with a schema based on TimeML[2] guidelines. We developed and evaluated approaches for: extraction of temporal expressions (TIMEX3) and EVENTS; TIMEX3 and EVENT attributes; document-time relations; and narrative container relations. Our approach is based on supervised learning (CRF and logistic regression), utilizing various sets of syntactic, lexical and semantic features with addition of manually crafted rules. Our system demonstrated substantial improvements over the baselines in all the tasks.

1 Introduction

SemEval-2016 Task 12 (Clinical TempEval) is a direct successor to 2015 Clinical TempEval (Bethard et al., 2015) and the past I2b2 temporal challenge (Sun et al., 2013). Clinical TempEval is designed to address the challenge of understanding clinical timeline in medical narratives and it is based on the THYME corpus (Styler IV et al., 2014) which includes temporal annotations.

Researchers have explored ways to extract temporal information from clinical text. Velupillai et al. (2015) developed a pipeline based on ClearTK[3] and SVM with lexical features to extract TIMEX3 and EVENT mentions. In 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). Besides THYME corpus, there have been other efforts in clinical temporal annotation including 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 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), hierarchical topic modeling (Alfonseca et al., 2012), and SVM with Tree Kernels (Miller et al., 2013).

Clinical TempEval 2016 was focused on designing approaches for timeline extraction in the clinical domain. There were 6 different tasks in the TempEval 2016, which are listed in Table 1. Per TimeML specifications (Pustejovsky et al., 2003), we refer to temporal expressions as TIMEX3 and events as EVENT throughout the paper. Attributes of TIMEX3 and EVENTS are outlined according to the THYME annotations (Styler IV et al., 2014). 16 teams participated in TempEval 2016 (Bethard et al., 2016).

For extracting temporal information from clinical text, we utilize supervised learning algorithms

[1] Temporal Histories of Your Medical Event. https://clear.colorado.edu/TemporalWiki/index.php/Main_Page

[2] TimeML is a standard specification language for events and temporal expressions in natural language. http://www.timeml.org/

Proceedings of SemEval-2016, pages 1248–1255,
San Diego, California, June 16-17, 2016. ©2016 Association for Computational Linguistics

Task	Description
TS	TIMEX3 spans
ES	EVENT spans
TA Class	Attributes of TIMEX3 ⟨DATE, TIME, DURATION, QUANTIFIER, PREPOSTEXP, SET⟩
EA Modality Degree Polarity Type	Attributes of EVENTs ⟨ACTUAL, HYPOTHETICAL, HEDGED, GENERIC⟩ ⟨MOST, LITTLE, N/A⟩ ⟨POS, NEG⟩ ⟨ASPECTUAL, EVIDENTIAL, N/A⟩
DR	Relation between EVENT and document time ⟨BEFORE, OVERLAP, BEFORE/OVERLAP, AFTER⟩
CR	Narrative container relations

Table 1: Tasks of clinical TempEval 2016

(Conditional Random Fields (CRF) and logistic regression) with diverse sets of features for each task. We also utilize manually-crafted rules to improve the performance of the classifiers, when appropriate. We show the effectiveness of the designed features and the rules for different tasks. Our system outperforms the baselines across all tasks, and is above the median results of all the teams in all tasks but one (CR in precision)[1].

2 Methodology

Our approach to all tasks is based on supervised learning using lexical, syntactic and semantic features extracted from the clinical text. We also designed custom rules for some tasks when appropriate. Details are outlined below:

2.1 TIMEX3 and EVENT Span Detection (TS, ES)

To extract TIMEX3 and EVENT spans (TS and ES), we use a combination of linear-chain CRFs (Lafferty et al., 2001) with manually-crafted rules[1]. Linear-chain CRFs are one of the most robust structured prediction approaches in natural language processing. We train the CRF for detecting TIMEX3s and EVENTs using BIO (Begin Inside Outside) labeling. That is, for the TIMEX3 classifier, after tokenizing the text, each token is labeled as either "O," "B-TIMEX3," or "I-TIMEX3". Similarly, the event classifier labels the tokens as either "O" or "B-EVENT," as virtually all EVENT annotations are only one token long. We use the CRF-Suite toolkit (Okazaki, 2007) for our experiments.

Features	lowercase; token letter case; if token is title; if token is numeric; if token is stopword; POS tag; brown cluster; prefix; suffix; noun chunk shape of the token; lemma

Table 2: Base feature set for supervised algorithms

The main features that we use for CRF in TS and ES tasks are outlined in Table 2. Among these features is Brown clustering (Brown et al., 1992) which is a form of hierarchical clustering based on the contexts in which the words appear. Brown clusters mitigate lexical sparsity issues by considering the words in their related cluster. We constructed fifty clusters across the the train and test datasets and passed the binary identifier of a token's cluster as the feature.

In addition to these features, we use domain specific features for EVENT span detection. Our domain feature extraction is based on the Unified Medical Language System (UMLS) ontology (Bodenreider, 2004). We use MetaMap[2] (Aronson and Lang, 2010), a tool for mapping text to UMLS concepts, for extracting the concepts. The semantic types of the extracted concepts are then used as features. Since UMLS is very comprehensive, considering all the semantic types causes drift. Thus, we limit semantic types to those indicative of clinical events (e.g. diagnostic procedure, disease or syndrome, and therapeutic procedure). For each feature set, we expand the features by considering a context window of +/- 3 tokens (The context window of size 3 yielded the best results on the development set).

For EVENT spans, we supplement the CRF output spans with manually crafted rules designed to capture EVENT spans. Particularly, we add rules to automatically identify EVENTs relating to standard patient readings. For example in: *"Diastolic=55 mm[Hg]"*, using simple regular expressions, we isolate the word "Diastolic" as an EVENT span.

For TIMEX3 spans, we use regular expressions that were designed to capture standard formatted

[1] The official ranking of participating teams is unknown at the time of writing and would be announced at the SemEval workshop.

[2] https://metamap.nlm.nih.gov/; we used the 2014 version

	Features
Set 1	UMLS semantic type; tense of the related verb in dependency tree; dependency root of the sentence
Set 2	class, text and brown cluster of closest DATE, PREPOSTEXP and TIMEX3; comparison with section time; comparison with document time; sentence tense and modals

Table 3: Additional feature sets used for document-time relation (DR) extraction.

dates. These rules improved the results of ES and TS considerably, as shown in Section 3.2.1.

2.2 Time and Event Attribute Detection

The main attribute for TIMEX3 mentions is their "class," which can be one of the following six types: DURATION, DATE, QUANTIFIER, PREPOSTEXP, SET, or TIME. EVENTs have four attributes each of which includes different types (Table 1). Full description of the types of the attributes are described by Styler IV et al. (2014). To properly classify each TIMEX3 and EVENT attribute, we train a separate logistic regression classifier[1] for each TIMEX3 and EVENT attribute value. These classifiers are trained on the TIMEX3 and EVENT spans that were previously extracted from the span CRFs (Section 2.1), and employ a similar feature set as the others.

In addition to the base feature set, we also incorporate rules as features in our classifier. We consider words that are indicative of certain EVENT attributes. For example, words such as "complete" or "mostly" indicate DEGREE:MOST, "possibly" indicates MODALITY:HEDGED and "never" shows POLARITY:NEG. We add such contextual features for DEGREE, MODALITY, and POLARITY.

In addition to the rules mentioned above, we further devise rules that lead to immediate classification as a specific class or attribute value. For example, TIMEX3 annotations in the format "[number] per [number]" are classified as SET automatically. We use the most probable predicted class as the final assigned label.

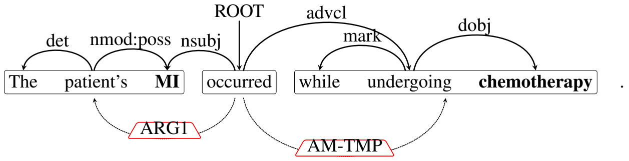

Figure 1: Example of dependency parse tree and semantic roles in the sentence. Dependency relations are shown by arrows above the sentence and semantic roles below it. The boldface terms in the sentence are event mentions. Per human annotation, the following relation exists in this sentence: *[MI]CONTAINS[chemotherapy]*.

2.3 Document-time Relation (DR)

Document-time relations (DR) are specific attributes of EVENTs indicating their temporal relation with the document creation time. There are 4 different types of DRs, namely, BEFORE, AFTER, OVERLAP, and BEFORE/OVERLAP. For identifying the DR attribute types, we use the same general classification approach as EVENT and TIMEX3 attributes; we train separate classifiers for each DR type using an extended set of features to what was used for EVENT attributes detection.

Table 3 describes the additional features that we use for DR extraction. In addition to the base features, we consider features specific to the EVENT annotation. These features are illustrated as Set 1 in table 3. We furthermore expanded the features by considering contextual features from the sentence and nearby time and date mentions (Set 2 in Table 3). Medical narratives often follow a chronological order. Therefore, nearest TIMEX3 mentions, and their comparison with the section timestamp or document timestamp can be good indicators of DRs. Similarly, verb tense and the modals in the sentence are also indicative of the sentence tense and can help in identifying the document-time relation. These additional features improved the results, as shown in Section 3.2.3.

2.4 Narrative Container Relations (CR)

Narrative containers (Pustejovsky and Stubbs, 2011) are TIMEX3s or EVENTs that subsume other EVENTs in a section of the text. They serve as temporal buckets into which other EVENTs fall. For example in the sentence: *"The patient recovered well after her first [surgery] on [December 16th]"*, [December 16th] is the narrative container and the

[1] We used the scikit-learn implementation. With L1 regularization and Liblinear solver.

Evaluation	Phase 1																							
Task	TS			ES			TA			EA												DR		
							CLASS			MODALITY			DEGREE			POLARITY			TYPE					
Metric	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baseline	.774	.428	.551	.878	.834	.855	.746	.413	.532	.810	.770	.789	.874	.831	.852	.812	.772	.792	.885	.813	.833	.620	.589	.604
Median	.779	.539	.637	.887	.846	.874	.755	.499	.618	.830	.780	.810	.882	.838	.869	.868	.813	.839	.854	.813	.844	.655	.624	.639
Our system	.802	.678	.735	.891	.872	.881	.775	.655	.710	.836	.818	.827	.887	.868	.877	.875	.856	.866	.868	.849	.858	.719	.704	.711

Table 4: Phase 1 evaluation results on test set for the tasks Time Span (TS), Event Span (ES), Time Attribute (TA), Event Attribute (EA) and Document-time Relation (DR). Baseline refers to the *memorize* baseline described in Section 3.1

.

following containment relation exists: "{December 16th} CONTAINS [surgery]".

To extract narrative container relations, we use the semantic frames of the sentences. We only consider the intra-sentence containment relations (appearing in the same sentence) and do not handle inter-sentence relations (crossing sentences). According to the THYME annotation guidelines (Styler IV et al., 2014), both EVENTs and TIMEX3s can provide boundaries of narrative containers. The first step in identifying narrative container relations is to identify the *anchor*, the EVENT or TIMEX3 span which contains all the other related EVENTs (*targets*). To learn the *anchor*, *target* and containment relation, in addition to the base features for *anchor* and *target*, we use Semantic Role Labeling (SRL) and dependency parse tree features of the sentence.

SRL assigns semantic roles to different syntactic parts of the sentence. Specifically, according to PropBank guidelines (Palmer et al., 2005), SRL identifies the semantic arguments (or predicates) in a sentence. If the *anchor* or the *target* fall in a semantic argument of the sentence, we assign the argument label as the feature to the associated *anchor* or the *target*. Using semantic roles, we extract the semantics of constituent parts of the sentence in terms of features which help to identify the container relations. For SRL, we use Collobert et al. (2011) neural model[1]. An example of semantic role labels is outlined in figure 1, in which labels below the sentence indicate the semantic labels.

Next, we consider the dependency parse tree of the sentence. Given the *anchor* and the *target* we traverse the dependency parse tree of the sentence to identify if they are related through a same root. In the sample sentence shown in figure 1, chemotherapy is the *anchor* and MI is another event which is the *target*. As shown, they are connected through the root of the sentence ("occurred").

Per annotation guidelines, TIMEX3 spans should receive higher priority over EVENTs for being labeled as the *anchor*. Therefore, we also consider the type of the expression (TIMEX3 or EVENT) as feature. Additional features such as UMLS semantic types, POS tags, dependency relations and verb tense of the sentence's root are also considered. To extract POS, syntactic and dependency-based features, we use the Spacy toolkit (Honnibal and Johnson, 2015).

3 Experiments

The 2016 Clinical TempEval task consisted of two evaluation phases. In phase 1, only the plain text was given and the TIMEX3 and EVENT mentions were unknown. In phase 2, which was only for DR and CR tasks, the TIMEX3 and EVENT mentions were revealed. In phase 1, we participated in all tasks, except for CR. In phase 2, we participated in both the DR and CR tasks.

3.1 Baselines

The baselines are two rule-based systems (Bethard et al., 2015) that are provided 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 match* baseline, predicts a time expression to be narrative container, if it is the closest EVENT expression.

3.2 Results

Our system's results on test set for all tasks are presented in Table 4 (phase 1) and Table 5 (phase 2).

[1] SENNA implementation: http://ml.nec-labs.com/senna/

Evaluation phase	Phase 2					
Task	DR			CR		
Metric	P	R	F1	P	R	F1
Baseline	-	.675	-	.459	.154	.231
Median	-	.724	-	.589	.345	.449
Our results	.816	.813	.815	.546	.471	.506

Table 5: Phase 2 evaluation results for Document-time Relation (DR) and narrative Containment Relations (CR) (The values indicated by (-) were not reported in SemEval official results). Baseline for DR is the *memorize* baseline and for CR is the *closest match* baseline (Section 3.1).

Model	P	R	F1
CRF	.758	.616	.679
CRF + rules	.777	.654	.710

Table 6: Effect of manually-crafted rules for TIMEX3 span (TS) on development set.

Our results in all tasks outperform the baselines, and in all but one case (CR-Precision) are above the median of all the participating teams.

3.2.1 TIMEX3 and EVENT spans (TS, ES)

For TS and ES, our system achieved F1 scores of 0.735 and 0.881 (on the test set) which gives +33.4% and +3.0% improvement over the baseline. While the improvement for TS is much larger, we observed less improvement on the ES task. For ES, Table 6 shows the effect of incorporating manually crafted rules to the output of CRF. These rules improved the F1 performance by 4.6%. In addition, as illustrated in Table 7, adding domain specific features (UMLS semantic types) improved the performance of base features (+2% F1). Adding manual rules to the output of CRF resulted in further improvement (additional +1% F1).

3.2.2 TIMEX3 and EVENT attributes (TA, EA)

For TA and EA, our system achieved an F1 of 0.710 and an average F1 of 0.856, respectively (Table 4). Our results improve over the baseline by 33.5% in TA and 4.8% in EA, respectively (for baseline, the average F1 of EA over all attribute types is 0.817). For EA, while performance of all types of attributes is comparable, the best performance relates to DEGREE attribute class with F1 of .887. The results of the TA and EA tasks for development set are also reported in Table 8. Generally, our results on the test set are marginally higher than on the development set which shows that we have successfully avoided over-fitting on the training and development sets.

Model	P	R	F1
base features	.863	.836	.849
+ UMLS	.879	.854	.866
+ rules	.886	.864	.875

Table 7: EVENT span (ES) results on development set based on different features.

Category	P	R	F1
TA:CLASS	.752	.632	.687
EA:MODALITY	.832	.816	.824
EA:DEGREE	.879	.863	.871
EA:POLARITY	.864	.848	.856
EA:TYPE	.854	.838	.846

Table 8: Results of the TIMEX3 and EVENT attributes (TA and EA) on the development set.

3.2.3 Document-time relation (DR)

The DR task was included in both evaluation phases. Its F1 score in phase 1 was 0.711 and in phase 2 was 0.815. Naturally, since in phase 1, the spans of EVENTs were unknown, lower performance is expected in comparison with phase 2. The DR results in both phases show substantial improvements over the baseline (+17.7% F1 in phase 1 and +20.4% recall in phase 2).

The effect of context window size on DR performance on development set is reported in Table 9. As the window size increases, more contextual features are added and therefore performance increases. However, after a certain point, when the window becomes excessively large, the performance decreases. We attribute this to overfitting the training data because of too many features. The optimal context window size is 6 which we used for our final submission. As far as features, we evaluated three primary feature sets (using a window of 6), the results of which are outlined in Table 10. The features are defined in tables 2 and 3. As illustrated, the addition of Set 1 and Set 2 features resulted in improvements in all DR types.

Error analysis for DR showed that many of the misclassified examples were for the BEFORE/OVERLAP relations, as also reflected in the low relative performance of BEFORE/OVERLAP relations (Table 10). In many cases, these relations are wrongly classified as either BEFORE or OVERLAP categories. For some cases, it is

w (+/-)	P	R	F1
1	.781	.780	.781
2	.790	.789	.790
3	.794	.793	.794
4	.799	.798	.798
5	.801	.800	.801
6	.804	.802	.803
7	.802	.801	.802
8	.795	.793	.794

Table 9: DR results on development set by window size.

not clear even for human whether the EVENTs had happened before the creation time of the document or they continued at document creation time. For example in the following: "*Resected rectal [adenocarcinoma], with biopsy-proven local [recurrence]*", the EVENT [adenocarcinoma] is of document relation type BEFORE/OVERLAP whereas our classifier wrongly classified them as BEFORE.

3.2.4 Narrative Container relations (CR)

The CR results are presented in Table 5. Our approach substantially improves over the baseline, especially in terms of recall (+2.06 times recall improvement). This demonstrates that using semantic frames of the sentences as well as their dependency structure can be effective in identifying container relations. However, F1 score of 0.506 shows that there is still plenty of room for improvement on this task. Error analysis showed that many of the false negatives relate to the inter-sentence relations. Our approach is designed for capturing only intra-sentence container relations. Similarly, some other false negatives were due to the dates that were not syntactically part of the sentence. An example is: "{*June 14, 2010*}*: His first [colonoscopy] was positive for [polyp]*". In this example, {June 14, 2010} is the anchor and *[colonoscopy]* and *[polyp]* are the targets. However, the designated date is not any syntactic part of the sentence and consequently, our approach is unable to capture that as the correct anchor of the narrative container.

4 Discussion and conclusions

Features	DR Type				
	All	After	Bef.	B/O	Over.
Base	.785	.725	.791	.536	.820
+ Set 1	.792	.751	.796	.545	.823
+ Set 2	.803	.756	.812	.538	.833

Table 10: DR Results breakdown by type based on different features on the development set. Base features are defined in Table 2, Set 1 and 2 features are defined in Table 3.

SemEval 2016 task 12 (Clinical TempEval) was focused on temporal information extraction from clinical narratives. We developed and evaluated a system for identifying TIMEX3 and EVENT spans, TIMEX3 and EVENT attributes, document-time relations, and narrative container relations. Our system employed machine learning classification scheme for all the tasks based on various sets of syntactic, lexical, and semantic features. In all tasks, we showed improvement over the baseline and, in all but one case (CR-Precision) we placed above the median of all participants (The official ranking of the systems were not announced at the time of writing).

While we showed the effectiveness of diverse set of features along with supervised classifiers, we also illustrated that incorporating manually crafted extraction rules improves results. However, manual rules should be constrained as some rules interfere with the learning algorithm and negatively affect the results. The strongest rules were those based on consistent patterns, such as dates in the standard format (e.g. MM-DD-YYYY). On the other hand, while some other rules improved the recall, they led to much lower precision and F1 score. For example, a rule that matches the word "time" as TIMEX3 span, improved our TS recall considerably but at the expense of overall precision and therefore was not included in the final submission.

For narrative containment relations, we showed that semantic frames and dependency structure of the sentence are helpful in identifying the relations. However, our approach is limited to intra-sentence relations and we are not detecting relations that are cross-sentences. In future work, we aim to expand our approach to detect inter-sentence container relationships.

Acknowledgments

We thank the reviewers for their comments. This work was partially supported by National Science Foundation through grant CNS-1204347 and REU award IIP-1362046.

References

Enrique Alfonseca, Katja Filippova, Jean-Yves Delort, and Guillermo Garrido. 2012. Pattern learning for relation extraction with a hierarchical topic model. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 54–59. Association for Computational Linguistics.

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Steven Bethard, Leon Derczynski, Guergana Savova, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. Semeval-2015 task 6: Clinical tempeval. *Proc. SemEval*.

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.

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*, pages 698–706. Association for Computational Linguistics.

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*, pages 173–176. Association for Computational Linguistics.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

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*, pages 715–720. ACM.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, September. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Timothy Miller, Steven Bethard, Dmitriy Dligach, Sameer Pradhan, Chen Lin, and Guergana Savova, 2013. *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, chapter Discovering Temporal Narrative Containers in Clinical Text, pages 18–26. Association for Computational Linguistics.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

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 and Amber Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 152–160. Association for Computational Linguistics.

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 Ward. 2009. Towards temporal relation discovery from the clinical narrative. In *AMIA*.

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).

Sumithra Velupillai, Danielle L Mowery, Samir Abdelrahman, Lee Christensen, and Wendy W Chapman. 2015. Blulab: Temporal information extraction for the 2015 clinical tempeval challenge. Association for Computational Linguistics.

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*, pages 405–413. Association for Computational Linguistics.