UtahBMI at SemEval-2016 Task 12: Extracting Temporal Information from Clinical Text

Abdulrahman Khalifa
University of Utah
abdulrahman.aal@utah.edu

Sumithra Velupillai
KTH, Stockholm/King's College, London
sumithra@kth.se

Stephane Meystre
University of Utah
stephane.meystre@hsc.utah.edu

Abstract

The 2016 Clinical TempEval continued the 2015 shared task on temporal information extraction with a new evaluation test set. Our team, UtahBMI, participated in all subtasks using machine learning approaches with ClearTK (LIBLINEAR), CRF++ and CRFsuite packages. Our experiments show that CRF-based classifiers yield, in general, higher recall for multi-word spans, while SVM-based classifiers are better at predicting correct attributes of TIMEX3. In addition, we show that an ensemble-based approach for TIMEX3 could yield improved results. Our team achieved competitive results in each subtask with an F1 75.4% for TIMEX3, F1 89.2% for EVENT, F1 84.4% for event relations with document time (DocTimeRel), and F1 51.1% for narrative container (CONTAINS) relations.

1 Introduction

Extracting temporal information from unstructured clinical narratives is an important step towards the accurate construction of a patient timeline over the course of clinical care (Savova et al., 2009), identifying and tracking patterns of care that are crucial for decision making (Augusto, 2005; Wang et al., 2008) and identifying cases or cohorts with temporal criteria for medical research (Raghavan et al., 2014). In the medical domain, more emphasis has been placed on utilizing temporal information from structured databases (Combi et al., 2010). However, recent developments in Medical Natural Language Processing (NLP) research has stimulated work in extracting information from unstructured clinical text (Meystre et al., 2008; Velupillai et al., 2015a) and facilitated future directions to extracting temporal information (Zhou and Hripcsak, 2007).

The i2b2 series of NLP challenges focused in 2012 on extracting events (problems, treatments and tests), time expressions (date, duration, time and frequency) and temporal relations (before, after, overlap) from a set of annotated discharge summaries. The best performing systems used supervised machine learning approaches, except for time expression identification and normalization where rule-based followed by hybrid approaches were most successful (Sun et al., 2013b; Sun et al., 2013a).

In 2015, the SemEval challenge included a Clinical TempEval task (Bethard et al., 2015) with similar objectives to the 2012 i2b2 challenge. The TimeML event and temporal expressions specification language (Pustejovsky et al., 2010) was adapted to define events, time expressions and relation annotations suitable for the clinical domain (Styler et al., 2014). The THYME (Temporal Histories of Your Medical Event) corpus is used in the Clinical TempEval challenge. The annotations in this corpus introduce the use of narrative containers concept (Pustejovsky and Stubbs, 2011) to reduce the complexity of finding temporal relations between every possible pair, and allow rapid discovery through automatic inferences. Each event and time expression is, when possible, assigned a narrative container that defines their temporal span. Groups of events and times within a narrative container can then be linked as one unit with other containers; eliminating the need to explicitly link every pair of events and times.

Proceedings of SemEval-2016, pages 1147–1153,
San Diego, California, June 16-17, 2016.

The additional pairs can be derived easily from minimal links between pairs within different narrative containers.

We present in this paper the methods used and results obtained from experiments with SVM-based linear classifiers and CRF-based sequential classifiers for the Clinical TempEval task. We complement the paper with a discussion and insights that potentially could help future efforts in this domain.

2 Methods

2.1 Task & Materials

The 2016 Clinical TempEval challenge included 6 subtasks: TIMEX3 1) span detection and 2) attribute classification, EVENT 3) span detection and 4) attribute classification, 5) relation between each event and document creation time classification (known as DocTimeRel), and narrative container or 6) CONTAINS relations between pairs of events and times classification. Our team participated in both phases provided in the challenge (phase 1: plain text only of the test set and phase 2: reference annotations for TIMEX3 and EVENTS including attributes were given for the relation classification subtasks) For a detailed description of the subtasks and evaluation metrics we refer the reader to (Bethard et al., 2015; Bethard et al., 2016).

The THYME corpus used in this task consists of treatment and pathology notes for colon cancer patients from the Mayo clinic. Three datasets were provided: *train* (=293 documents), *dev* (=147) and *test* (=151). We used the dev set to benchmark different approaches during system development and as a guideline to manually select the best performing features. All final models used for predictions were trained using the combined *train+dev* datasets. The test set was used for the final evaluation. Each subtask was addressed separately using a machine learning classifier and groups of almost similar features with slight changes such as surrounding context window sizes. cTAKES (Savova et al., 2010) was used to pre-process each clinical note to generate morphological, lexical and syntactic-level annotations, which were used as features for training the classifiers. The ClearTK machine learning package (Bethard et al., 2014) was used to build Support Vector Machine (SVM) LIBLINEAR (Fan et al., 2008) classifiers, while CRFsuite (Okazaki, 2007) and CRF++ (Kudo, 2005) were used to build Conditional Random Field (CRF) sequential classifiers. Both cTAKES and ClearTK utilize the Apache Unstructured Information Management Applications (UIMA) framework (Ferrucci and Lally, 2004) which makes it easy to integrate modules from both applications and pipeline output from cTAKES to ClearTK using the XML Metadata Interchange (XMI) format.

2.2 Input Preparation/Feature Extraction

Each clinical note in the corpus was previously segmented into sections with a `[start section id=...]` and `[end section id=...]` markers that were easy to identify and annotate using regular expressions. Therefore, we built a UIMA module to segment each clinical note into section boundaries; each annotated with their respective section ID. cTAKES clinical pipeline (version 3.2.2) was used to extract lexical and syntactic features. These include sentence boundaries, tokens, lemmas, part-of-speech tags, syntactic chunk tags (e.g. Verb Phrase-VP, Noun Phrase-NP), token type as defined by cTAKES (see figure 1), as well as dependency parse and semantic role labels used for relation classification. Furthermore, ClearTK feature extractors were used to generate word shape features (e.g. capital, lower, numeric), character patterns and character N-gram features for the linear classifiers. The CRFsuite package comes with built-in feature extractor functions for word shapes, character patterns and N-gram which were used for the TIMEX3, EVENT and DocTimeRel CRF classifiers. Table 1 outlines the features used in each subtask.

For the CRF packages, the features had to be transformed into a flat, tab-separated structure with columns of tokens and associated features each placed in one line. Sentences are designated by empty lines following a sequence of lines of tokens (see Figure 1 for an example).

2.3 SVM-based Approach

The LIBLINEAR package within ClearTK was used to train all linear classifiers with default settings (C=1.0; s=1; Loss=dual L2-regularized) except for TIMEX3 (grid search performed on the training set indicated a better value for C=0.5). We re-used

Feature Type	CRFsuite TIMEX3	CRFsuite EVENT	CRFsuite DocTimeRel	CRF++ DocTimeRel	LIBLINEAR TIMEX3	LIBLINEAR EVENT	LIBLINEAR DocTimeRel	LIBLINEAR CONTAINS
Window Size (preceding, following)	$-2, +2$	$-2, +2$	$-2, +2$	$-5, +5$	$-5, +5$	$-2, +2$	$-5, +5$	$-5, +5$
Token	*	*	*	*	*	*	*	*
Token (lowercased)	*	*	*		*	*		
Lemma	*	*	*	*			*	
Part of Speech (POS)	*	*	*	*	*	*	*	*
Chunk Type	*	*	*	*				
Token Type (WORD, NUMERIC, ...)	*	*	*	*	*			
Word Shape (ALL-CAP, INITIAL-CAP, ...)	*	*	*		*			
Section ID	*	*	*	*	*	*	*	*
Character Pattern	*	*	*		*			
Character Ngram	*	*	*		*			
EVENT and attributes Tags							*	*
TIMEX3 and attributes Tags							*	*
HeidelTime Token					*			
TIMEX position in sentence								*
Number of tokens between relation pair								*
Semantic role arguments								*
C Parameter	1.0	1.0	1.0	1.0	0.5	1.0	1.0	1.0

Table 1: List of features used (indicated with asterisk) for each subtask with different machine learning approaches.

```
token    lemma    pos chunk   token_type      section_ID  ...
#        #        NN  B-NP    SymbolToken     20112       ...
1        1        LS  I-NP    NumToken        20112       ...
Dilated dilat     JJ  I-NP    WordToken       20112       ...
```

Figure 1: Example of the flat input used for the CRF approaches: features in columns separated by tabs.

the approach taken in the 2015 Clinical TempEval (Velupillai et al., 2015c) for TIMEX3, EVENT and DocTimeRel subtasks, with minor changes in the used features. For TIMEX3, one separate classifier was created for each class (e.g., DATE, TIME). For EVENT, one classifier was created for detecting the text span, and one separate classifier for each attribute (i.e., MODALITY, DEGREE, POLARITY and TYPE). In addition, we added a classifier in this pipeline, for event relations with the document time (DocTimeRel). The main feature additions in this year's challenge were a section ID feature for all classifiers; and a binary feature— whether or not a token was classified as temporal expression of an adapted version of HeidelTime (Strötgen and Gertz, 2010) — for the TIMEX3 subtask.

For the narrative container (CONTAINS) relations subtask, we trained four models to predict relations between pairs of 1) event-event and 2) event-time within a sentence; and 3) event-event and 4) event-time across consecutive sentences. This approach has been previously shown to be most effective in predicting temporal relations (Xu et al., 2013). The candidate pairs were selected[1] using the following strategy: All possible combinations of pairs between events and events-times within a sentence were generated for training and classification. For event-event pairs across consecutive sentences; only the first and last event from the current sentence were paired with the first and last from next (or subsequent) sentence. For event-time pairs across sentences; each time phrase in the current sentence is paired with the first and last events from the preceding and following sentences. This approach suffers from the limitation of allowing many examples with the negative class (i.e., pairs without a relation) to be selected; and hence causes class imbalance that may affect classifier training. (Tang et al., 2013) demonstrated that using heuristics to select candidates that are more likely to be part of a relation could produce superior results for temporal relation classification. Another possible remedy is to introduce scaling parameters to adjust the weight of each class during training, such that data samples from the positive class get more weight while the negative class samples get less weight (Lin et al., 2015). Due to time constraints, we were unable to experiment with either of these approaches.

2.4 CRF-based Approach

For the sequential classification, we used the CRFsuite for TIMEX3, EVENT and DocTimeRel subtasks in phase 1, and CRF++ for the DocTimeRel subtask in phase 2. All CRF trained models used default settings (C=1.0; algorithm=L-BFGS). During phase 1, we employed a cascaded approach: we trained CRFsuite models to 1) predict textual spans of TIMEX3 and EVENT tokens separately; 2) pre-

[1] cTAKES Temporal module was very useful in facilitating experiments for the TLINK relations.

	span			span+class		
	P	R	F1	P	R	F1
MAX	0.840	0.758	0.795	0.815	0.735	0.772
CRFsuite	0.798	0.714	0.754	0.771	0.690	0.729
LIBLINEAR	**0.810**	0.690	0.745	**0.792**	0.674	0.728
CRFsuite+LIBLINEAR	0.761	**0.769**	**0.765**	0.733	**0.741**	**0.737**
memorize (Baseline)	0.774	0.428	0.551	0.746	0.413	0.532

Table 2: TIMEX3 subtask results on the test set.

dict TIMEX3 and EVENT attributes using the predictions in step 1), and 3) predict DocTimeRel and CONTAINS relations using the predictions in steps 1-2. The prediction labels were encoded using the standard IOB2 format of **I**nside, **B**egin, and **O**utside. For instance, prediction labels for the phrase "see him this afternoon ." will be encoded as "O O B-TIME I-TIME O" where "this afternoon" is a TIMEX3 expression in this context. CRF classifiers are probabilistic graphical models that take into account a previous window of prediction labels and assign the most likely sequence of labels based on estimates obtained from the training data. Therefore, they usually perform better in tasks that require assigning labels to sequential data. This is particularly true for the TIMEX3 subtask where the majority of time phrases span multiple tokens.

3 Results

The performance we obtained for the various subtasks on the test set are shown in Tables 2, 3, 4. We also include the results from two baseline systems (**memorize** — for EVENT, TIMEX3 and DocTimeRel, and **closest** — for CONTAINS relations) provided by the workshop organizers, as well as the maximum score achieved in each subtask from all submissions (Bethard et al., 2016). Note that for the narrative container subtask, we report the official score and corrected score we obtained after discovering and correcting a bug affecting the LIBLINEAR models that prevented predictions of event-time relations.

CRF achieved a better performance (F1 %75.4) than the linear classifier (F1 %74.5) when detecting TIMEX3 spans because of higher recall (R %71.4). The LIBLINEAR model resulted in higher precision (P %81). Our initial analysis indicates that this is partly due to many CRF predictions overlapping with the reference annotations rather than matching exactly. When using a strict match evaluation approach, these overlaps are counted as false positives. For example, the CRF approach generated TIMEX3 labels for expressions like "at the time" and "in the past" while the reference standard included TIMEX3 annotations for only "the time" and "past", respectively. Combining the predictions from both models (by taking the union set of outputs and discarding duplicated predictions) allowed for improved performance (F1 %76.5) suggesting that an ensemble-based strategy could yield superior results for this subtask. Additional analysis will be needed to understand which class of TIMEX3 phrases each model is better at predicting and apply a more sophisticated ensemble method such as weighted average.

The results for the EVENT subtasks were almost identical between the two approaches (CRF or LIBLINEAR), except when classifying the *modality* and *type* attributes where CRF performed better. Combining the predictions from both models did not allow for any performance improvements. Note also that the baseline results for this subtask are very high.

For the DocTimeRel subtask, the CRFsuite model reached an F1 of %74.5 in phase 1, while the CRF++ model reached an F1 of %84.4 in phase 2; allowing for significant improvement over the performance of the LIBLINEAR model (F1 %81.8). For the CONTAINS relations classification subtask, the LIBLINEAR models achieved an F1 of %42.2 in phase 1 when using CRF predictions of TIMEX3 and EVENT; and F1 of %51.1 in phase 2. Note that for phase 2 we also included the prediction of DocTimeRel relations from CRF as an input feature to the LIBLINEAR models.

4 Discussion

Several important issues need to be addressed for future improvement in this task or other similar tasks. We outline some of these issues below, along with an analysis from the reference standard annotations and the system prediction errors.

The CRF-based classifiers detected TIMEX3 mentions with higher accuracy. As mentioned previously, many of these mentions were overlapping with the reference standard annotations. Our output included 352 false positive errors when using a strict match evaluation. Among these errors, about

	span			span+modality			span+degree			span+polarity			span+type		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
MAX	0.915	0.891	0.903	0.866	0.843	0.855	0.911	0.887	0.899	0.900	0.875	0.887	0.894	0.870	0.882
CRFsuite	0.902	0.883	0.892	0.850	0.832	0.841	0.898	0.879	0.889	0.885	0.867	0.876	0.875	0.857	0.866
LIBLINEAR	0.897	0.886	0.892	0.841	0.831	0.836	0.892	0.881	0.887	0.879	0.869	0.874	0.854	0.843	0.849
memorize (Baseline)	0.878	0.834	0.855	0.810	0.770	0.789	0.874	0.831	0.852	0.812	0.772	0.792	0.855	0.813	0.833

Table 3: EVENT subtask results on the test set.

	DocTimeRel			CONTAINS		
	P	R	F1	P	R	F1
Phase 1: End-to-End with plain text only						
MAX	0.766	0.746	0.756	0.531	0.471	0.479
CRFsuite	0.753	0.737	0.745	-	-	-
LIBLINEAR	0.741	0.732	0.736	0.553	0.341	0.422
LIBLINEAR†	-	-	-	0.502	0.215	0.301
memorize/closest (baseline)	0.620	0.589	0.604	0.403	0.067	0.115
Phase 2: Includes manual annotations of TIMEX3 and EVENT						
MAX	-	0.843	-	0.823	0.564	0.573
CRF++	0.844	0.843	0.844	-	-	-
LIBLINEAR	0.818	0.818	0.818	0.657	0.418	0.511
LIBLINEAR†	-	-	-	0.562	0.254	0.350
memorize/closest (baseline)	-	0.675	-	0.459	0.154	0.231

Table 4: Relation classification results on the test set. †Indicates official scores before bug correction.

228 were overlapping (but not matching perfectly) with reference annotations, and the remaining 124 errors were due to other reasons. If counting these overlapping errors as true positives instead of false positives, as in a partial match evaluation, significant accuracy improvements could be observed (P: 0.929, R: 0.833, F1: 0.878)[2]. Contributions from last year's TempEval task have pointed out the issue of TIMEX3 annotations inconsistency in the reference standard (Tissot et al., 2015). After examining the 228 overlapping false positive errors further, we noticed through empirical analysis that many were due to either missing or added prepositions (e.g., 'at', 'in', 'for', 'about') and determiners ('a', 'the'). Further examination revealed that, as pointed out by the previous authors, there is an inconsistent trend in the reference standard annotations. For example, the reference standard contains the following TIMEX3 phrases (underlined words indicate words not annotated in the reference standard): "in the past", "in the last three days", "for many years", "for two years", "at this time", "at this time", "about 27 years ago" and "about 30 years ago". These irregularities will make it difficult for any machine learning model to generalize well beyond the given dataset and most likely will indicate overfitting for higher performance models (Velupillai et al., 2015b). The reported inter-annotator agreement for TIMEX3 span annotations of F1 77.4% (Bethard et al., 2015) further supports these assumptions. Therefore, future work should focus on creative ways to deal with this inconsistency and enable more generalizable solutions. Apart from the overlapping errors due to reference standard inconsistencies; other types of errors may indicate room for future improvement. We believe that training multiple classifiers and combining the outputs using ensemble-based approach could yield superior results as manifested from combining predictions of CRF and LIBLINEAR models.

For the DocTimeRel subtask, the CRF-based classification approach also allowed for significant improvements, particularly in phase 2. Table 5 shows the confusion matrix and evaluation scores obtained on the dev set for each category of DocTimeRel relation using CRF++ model when trained on the training set. The final scores achieved (R 83.3%) on the dev set, are comparable to the scores achieved (R 84.3%) on the test set. This allows us to make consistent conclusions about classifier performance on one set (dev) that can be expected to apply on the other set (test). The lowest accuracy (R 48.6%) was observed with the BEFORE/OVERLAP category. A possible explanation for this lower accuracy is the small number of training samples available in this category (2160 instances in the training set out of 38885). The confusion matrix shows that this category gets almost a balanced error rate between the BEFORE (297) and OVERLAP (271) categories. In addition, the highest number of misclassified instances occur in OVERLAP (972) and BEFORE (858) categories where one category is confused for the other. Future work should focus on improving classification in the BEFORE and OVERLAP categories.

The performance achieved using LIBLINEAR models in the CONTAINS relations subtask (F1 42.2%–51.1%) is a significant improvement over last year's attempt using a CRF model (F1 12.3%–

[2]This score was obtained using the `--overlap` option from the official evaluation script.

		SYSTEM				
		AFTER	BEFORE	BEFORE/OVERLAP	OVERLAP	TOTAL
REFERENCE	AFTER	**1686**	157	5	289	2137
	BEFORE	110	**6667**	145	972	7894
	BEFORE/OVERLAP	12	297	**548**	271	1128
	OVERLAP	231	858	145	**8579**	9813
	TOTAL	2039	7979	843	10111	20972
	SCORE (P/R/F1)	0.827/0.789/0.807	0.836/0.845/0.840	0.650/0.486/0.556	0.848/0.874/0.861	0.831/0.833/0.831

Table 5: Confusion matrix and scores for each category of DocTimeRel relation obtained on the dev set using CRF++ classifier.

26.0%) (Velupillai et al., 2015c). We think that studying different strategies for candidate pair selection or experimenting with different class weights to reduce effects of negative class predictions could allow for improvement in this subtask. In addition, although we used two separate models to predict relations between event pairs within and between consecutive sentences, we restricted the way we chose candidates across sentences (first and last from current sentence are paired with first and last from next sentence). This restriction was used to avoid an increase in the number of pairs without a relation (i.e., negative class pairs); in addition to the increased computational runtime penalty. However, this means that any candidate pairs spanning across many sentences will be missed by our classifier. This is especially true for some event and time phrases that are usually at the beginning of a sentence (mostly introducing a section header) and act as narrative containers for many events in the next few sentences. For instance, our classifier missed the 'HISTORY' narrative container appearing as part of the section header "PAST MEDICAL HISTORY", which is usually a relation source for many events discussed within the section. One example from the dev set shows that the 'HISTORY' event CONTAINS following events (e.g., medical conditions in a numbered list) spanning from the next first sentence down to the eleventh sentence. Future work could focus on using carefully hand-crafted rules to capture these pairs to increase recall. We think that the most successful approach for this subtask could use hybrid approaches combining rules and machine learning classifiers to improve recall and retain high precision, respectively.

5 Conclusion

Temporal information extraction and reasoning from clinical text remains a challenging task. Our analysis of different machine learning approaches have been informative, and resulted in competitive results for the 2016 Clinical TempEval subtasks. We plan to develop hybrid and ensemble-based approaches in the future to further improve performance on this, and other clinical corpora.

Acknowledgments

We would like to thank the Mayo clinic and the 2016 Clinical TempEval challenge organizers for providing access to the clinical corpus, and arranging NLP shared task.

References

Juan Carlos Augusto. 2005. Temporal reasoning for decision support in medicine. *Artificial intelligence in medicine*, 33(1):1–24, jan.

Steven Bethard, Philip V Ogren, and Lee Becker. 2014. Cleartk 2.0: Design patterns for machine learning in uima. In *LREC*, pages 3289–3293.

Steven Bethard, Leon Derczynski, Guergana Savova, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. SemEval-2015 Task 6 : Clinical TempEval. pages 806–814.

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.

Carlo Combi, Elpida Keravnou-Papailiou, and Yuval Shahar. 2010. *Temporal information systems in medicine*. Springer Science & Business Media.

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*, 9(2008):1871–1874.

David Ferrucci and Adam Lally. 2004. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

Taku Kudo. 2005. Crf++: Yet another crf toolkit (2005). *Software available at http://crfpp. sourceforge. net*. Accessed: 2010-02-25.

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

S M Meystre, G K Savova, K C Kipper-Schuler, and J F Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, pages 128–44, jan.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs), 2007. *URL http://www. chokkan. org/software/crfsuite*. Accessed: 2016-02-25.

James Pustejovsky and Amber Stubbs. 2011. Increasing Informativeness in Temporal Annotation. *Law*, (June):23–24.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An International Standard for Semantic Annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Preethi Raghavan, James L Chen, Eric Fosler-Lussier, and Albert M Lai. 2014. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? *AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science*, 2014:218–23.

Guergana Savova, Steven Bethard, Will Styler, James Martin, Martha Palmer, James Masanz, and Wayne Ward. 2009. Towards temporal relation discovery from the clinical narrative. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2009:568–72, jan.

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

Jannik Strötgen and Michael Gertz. 2010. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324. Association for Computational Linguistics.

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

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013a. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association : JAMIA*, 20(5):806–13.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013b. Temporal reasoning over clinical text: the state of the art. *Journal of the American Medical Informatics Association : JAMIA*, 20(5):814–9.

Buzhou Tang, Yonghui Wu, Min Jiang, Yukun Chen, Joshua C Denny, and Hua Xu. 2013. A hybrid system for temporal information extraction from clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 20(5):828–35.

Hegler Tissot, Genevieve Gorrell, Angus Roberts, Leon Derczynski, Marcos Didonet, and Del Fabro. 2015. UFPRSheffield : Contrasting Rule-based and Support Vector Machine Approaches to Time Expression Identification in Clinical TempEval. *Proc. SemEval*, (SemEval):835–839.

Sumithra Velupillai, D Mowery, BR South, M Kvist, and H Dalianis. 2015a. Recent advances in clinical natural language processing in support of semantic analysis. *Yearbook of medical informatics*, 10(1):183.

Sumithra Velupillai, D. L. Mowery, S. Abdelrahman, L. Christensen, and W. W. Chapman. 2015b. Towards a Generalizable Time Expression Model for Temporal Reasoning in Clinical Notes. In *AMIA 2015 Proceedings*, pages 1252–1259, San Francisco, USA, November. American Medical Informatics Association.

Sumithra Velupillai, Danielle L Mowery, Samir Abdelrahman, Lee Christensen, and Wendy W Chapman. 2015c. BluLab : Temporal Information Extraction for the 2015 Clinical TempEval Challenge. *Proc. SemEval*, (SemEval):815–819.

Taowei David Wang, Catherine Plaisant, Alexander J Quinn, Roman Stanchak, and Shawn Murphy. 2008. Aligning Temporal Data by Sentinel Events : Discovering Patterns in Electronic Health Records. *CHI 2008 Proceedings Health and Wellness*, pages 457–466.

Yan Xu, Yining Wang, Tianren Liu, Junichi Tsujii, and Eric I-Chao Chang. 2013. An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association : JAMIA*, 20(5):849–58.

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