BluLab: Temporal Information Extraction for the 2015 Clinical TempEval Challenge

Sumithra Velupillai[1,2], **Danielle L Mowery**[2], **Samir Abdelrahman**[2], **Lee Christensen**[2] **and Wendy W Chapman**[2]
[1]Dept. of Computer and Systems Sciences (DSV)
Stockholm University, Sweden
[2]Department of Biomedical Informatics
University of Utah, Salt Lake City
sumithra@dsv.su.se, {firstname.lastname}@utah.edu

Abstract

The 2015 Clinical TempEval Challenge addressed the problem of temporal reasoning in the clinical domain by providing an annotated corpus of pathology and clinical notes related to colon cancer patients. The challenge consisted of six subtasks: TIMEX3 and event span detection, TIMEX3 and event attribute classification, document relation time and narrative container relation classification. Our BluLab team participated in all six subtasks. For the TIMEX3 and event subtasks, we developed a ClearTK support vector machine pipeline using mainly simple lexical features along with information from rule-based systems. For the relation subtasks, we employed a conditional random fields classification approach, with input from a rule-based system for the narrative container relation subtask. Our team ranked first for all TIMEX3 and event subtasks, as well as for the document relation subtask.

1 Introduction

Temporal information extraction plays a crucial role in improved information access, in particular for creating timelines and detailed question answering. Several previous natural language processing (NLP) research community challenges have dealt with temporal reasoning in the newswire domain (Verhagen et al., 2010; UzZaman et al., 2013) and the clinical domain (Sun et al., 2013).

The 2015 Clinical TempEval challenge (Bethard et al., 2015) addressed temporal reasoning subtasks similar to these previous efforts by providing a new benchmark corpus in the clinical domain with annotated pathology and clinical notes from colon cancer patients. The corpus is annotated with a modified version of the TimeML schema (Pustejovsky et al., 2010), where adaptations specific to this domain have been developed (Styler et al., 2014).

For successful temporal modelling, three core concepts need to be defined: **temporal expressions (TIMEX3)**, *denoting time references like dates*; **events (EVENT)**, *denoting salient occurrences*; and **temporal relations (TLINK)** *denoting order (e.g. before, after) between an event and/or TIMEX3*.

As part of the 2012 i2b2/VA Challenge, the best performing systems for classification of TIMEX3 (F1: 0.66), EVENTS (F1: 0.92), their attributes (average accuracy: 0.86) and TLINKS (F1: 0.69) applied regular expressions as well as machine learning approaches such as conditional random fields (CRF) and support vector machines (SVM) (Sun et al., 2013). For the 2013/2014 CLEF/ShARe Challenges, the best approaches for strict information extraction (F1: detection and accuracy: normalization) of TIMEXs (0.287 F1 and 0.354 accuracy), disease/disorder EVENTS (0.750 F1 and 0.589 accuracy), and EVENT attributes (0.676 F1 and 0.868 accuracy) leveraged the Apache cTAKES (Savova et al., 2010) framework, Begin-Inside-Outside (BIO) tagging, and CRF and SVM for (Pradhan et al., 2015; Mowery et al., 2014).

The 2015 Clinical TempEval consisted of six subtasks related to these core concepts: TIMEX3 span (TS) and attribute (TA) classification, EVENT span (ES) and attribute (EA) classification, document creation time (DR) and narrative container (CR) rela-

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 815–819,
Denver, Colorado, June 4-5, 2015.

tions. Our team participated in all six subtasks, with the aim of benchmarking existing tools and methods on this corpus for further development of semantic processing of clinical notes. In this paper, we describe our system, its results, and an error analysis for each of the challenge subtasks.

2 Methods

We received 293 training reports for system development and 147 testing reports for blind system evaluation. For all subtasks, we extracted morphological (lemma), lexical (tokens), and syntactic (part-of-speech) features encoded from cTAKES. In the following sections, we enumerate additional subtask-specific features from various NLP systems used to train supervised learning (combined with rule-based in some cases) approaches for each subtask.

2.1 TIMEX3, EVENTS, and their Attributes

A UIMA pipeline using ClearTK (Bethard et al., 2014) was built for the subtasks TS, TA, ES and EA, using SVM classifiers (Liblinear) with parameters (C-value) set manually using a grid search. For TS, a separate classifier was built for each TA type using simple lexical features (the token itself in full and without its ending (2 characters), part-of-speech tag, numeric type, capital type, lower case, surrounding tokens) and gazetteer information based partly on an adapted version of HeidelTime (Strötgen and Gertz, 2013). Each token was classified as either B (Begin), I (Inside) or O (Outside) using the ClearTK BIO-chunking representation. Slightly different context window sizes and gazetteer information were employed for each TA value. For ES, one classifier was built for classifying tokens using the same BIO-chunking representation, employing similar lexical features and a context window size of ± 2, as well as a chunk type feature, followed by separate classifiers for each EA value. The values for TA and EA can be found in Table 1.

For EA, we used lexical features (similar to those used for TS and ES) along with new features from the pyConText system (Chapman et al., 2011). For each non-default EA, we evaluated the predictiveness of each cue from the pyConText linguistic knowledge base on the training set to determine its association. For example, the "denies" predicts **polarity**: NEG. We eliminated cues that were not relevant for the task e.g., **experiencer**. We then conducted an error analysis on the training data for missed cues and added them to the existing knowledge base for final evaluation. These cues were provided to the SVM model in addition to section information and previous EA assignments for each ES. For TA and EA, we used adapted versions of pyConText and HeidelTime as baselines.

Attribute	Potential Values
TA: **type**	*DATE, TIME, DURATION, QUANTIFIER, PREPOSTEXP or SET
EA: **modality**	*ACTUAL, HEDGED, HYPOTHETICAL or GENERIC
EA: **degree**	*N/A, MOST or LITTLE
EA: **polarity**	*POS or NEG
EA: **type**	*N/A, ASPECTUAL or EVIDENTIAL

Table 1: Possible values for TIMEX3 attributes (TA) and event attributes (EA). *default majority value.

2.2 DocTimeRel and Contains Relations

The challenge relation classification task consisted of two subtasks: DocTimeRel (DR) and narrative container relation (CR). For DR, the task was defined to identify 4 classes: before, after, overlap, and before/overlap which describe the relation between the event mentioned in the document and the related document time. For CR, the task was defined for the contains class to recognize whether one event/time mention in the document contains or is contained by another.

We used token-level features for each sentence. We parsed the cTAKES output to extract the following features: a binary feature indicating if the token is the first token in the sentence, the token lemma and normalization forms, its type of token (word/punctuation/symbol/number/contraction) and if it was tagged as any of the following semantic types by cTAKES: medical, procedure, anatomical site, sign/symptom, disease/disorder, and concept. We also added a feature indicating whether the token was part of an event mention, a time mention, or

none of these, extracted from the predictions (phase 1 in the challenge) or the gold annotations (phase 2).

We used CRF++[1] for the DR task using the aforementioned features along with a window of ±5 tokens for each feature as contextual features. For the CR task, we aimed at integrating machine learning (ML) and rule-based techniques as a potential solution. The search space was limited to three event or time mentions in ascending sequential order from the text to classify CR between two mentions. We used CRF++ again for the machine learning part, with the same token features as for DR. If two adjacent mentions were located in separate sentences, we merged the sentences to one.

For the rule-based part, we used the Moonstone system. Moonstone is a language processing tool which uses both a semantic grammar, and a rule engine which can take as input (among other things) the output of its grammatical parser (Christensen and Chapman, 2015). We situated Moonstone in a UIMA pipeline, along with the ClearTK predictions for TS, TA, ES, and EA, to recognize potential instances of the contains relation, using two rules which can be paraphrased in English as follows:

- If a DATE annotation initiates a sentence, and an EVENT annotation occurs anywhere in the following three sentences, with no intervening DATE mention, then infer a CR between the two.
- If two EVENT annotations appear within a sentence, and one appears commonly as the first argument in the training annotations denoting the contains relation, and the second commonly appears as the second contains argument in the training annotations, then infer a CR between the two.

Finally, to integrate both techniques, we conducted three runs. The first run (V1) was based entirely on the ML solution. In the second run (V2), we added the mentions extracted from the Moonstone rules to the V1 search space. In the third run (V3), we started with the mentions extracted from the Moonstone rules as an initial search space, then, we added pairs randomly from the first run such that each mention had maximum 3 nearest mentions including those of the Moonstone rules (if any).

[1]http://crfpp.googlecode.com/svn/trunk/doc/index.html, accessed Jan. 26 2015

Subtask	P	R	F1
TS	0.788	0.669	0.724
TS (b)	0.549	0.654	0.597
TA: **type**	0.772	0.658	0.710
TA (b): **type**	0.549	0.654	0.597
ES (*)	0.886	0.867	0.876
EA: **modality**	0.883	0.872	0.877
EA (b):**modality**	0.744	0.734	0.739
EA: **degree**	0.946	0.933	0.940
EA (b): **degree**	0.854	0.842	0.848
EA: **polarity**	0.931	0.919	0.925
EA (b): **polarity**	0.930	0.917	0.923
EA: **type**	0.894	0.883	0.888
EA (b): **type**	0.814	0.803	0.809

Table 2: Training set results for TIMEX3 span (TS), and attributes (TA), event span (ES), and attributes (EA). (b) = baseline. (*) For ES, no rule-based method was used as baseline, only different feature settings in ClearTK.

Subtask	P	R	F1
TS	0.797	0.664	0.725
TA: **type**	0.778	0.652	0.709
ES	0.887	0.864	0.875
EA: **modality**	0.834	0.813	0.824
EA: **degree**	0.882	0.859	0.870
EA: **polarity**	0.868	0.846	0.857
EA: **type**	0.834	0.812	0.823

Table 3: Test set results for TIMEX3 spans (TS), attributes (TA), event spans (ES), and attributes (EA).

3 Results

We present results on the training data and the final results on the test set for all challenge subtasks.

In Table 2, results on the training data for the TIMEX3 (TS, TA) and EVENT (ES, EA) tasks are shown, for the final ClearTK models that were used for system submission, as well as baseline results using adapted versions of pyConText and HeidelTime. The ClearTK modules resulted in improved performance for all subtasks. Final results on the test set are shown in Table 3.

For the relation subtasks DocTimeRel (DR) and narrative containers (CR), results on the training data are shown in Tables 4 and 5. For testing, two phases were provided in the challenge: one where

Subtask	P	R	F1
DR: **before**	0.814	0.801	0.807
DR: **overlap**	0.836	0.818	0.827
DR: **before-overlap**	0.745	0.736	0.740
DR: **after**	0.808	0.796	0.802
Overall	0.801	0.788	0.794

Table 4: Results for all relation types (before, overlap, before-overlap, after) for Document relation time (DR) on the training data.

Subtask	P	R	F1
CR: V1	0.118	0.124	0.121
CR: V2	0.142	0.266	0.185
CR: V3	0.160	0.176	0.168

Table 5: Results for the Contains relation (CR) on the training data. V# indicates the run.

only plain text was given (#1), and one where gold TIMEX3 and event annotations were given (#2). For CR, final results were calculated with or without closure. In Table 6 final results on the two relation tasks are shown.

Phase	Subtask	P	R	F1
1	DR	0.712	0.693	0.702
	CR V1	0.100	0.099	0.100
	CR V2	0.094	0.179	0.123
	CR V3	0.090	0.103	0.096
2	DR	-	-	0.791
	CR V1	0.117	0.128	0.123
	CR V2	0.140	0.254	0.181
	CR V3	0.150	0.155	0.153

Table 6: Results for DocTimeRel (DR) and narrative container relations (CR) on the test set. During Phase 1, only text was provided, while in Phase 2 manual EVENT and TIMEX3 annotations were provided. V# indicates the run. Results for CR are reported with closure.

4 Discussion

Our team had the highest F1 on all TIMEX3, EVENT and DR subtasks in the 2015 Clinical TempEval challenge. Similar to other best performing systems in previous temporal modelling challenges, we applied CRF, SVM, and rule-based approaches, using mostly simple features.

We observed moderate recall for TS which can be attributed to missing words ("perioperative") and span errors (e.g. "early July" (gold) vs. "early July apparently" (system)). TA values with very few training examples (e.g. **type**: TIME) were difficult for both approaches, with the exception of PREPOSTEXP, which resulted in high F1 on the training data. For ES, spanning issues were not the source for errors as much as for TS. Most errors were due to previously unseen words or contexts. For different EA types, rare classes were problematic, e.g. **degree**: LITTLE and MOST, but also distinguishing subtle differences between **modality**: GENERIC, HEDGED, and HYPOTHETICAL values.

In the DR subtask, we achieved high precision, recall, and F1 using simple cTAKES features. Careful analysis of our outputs revealed that some events have similar features with different relation classes. Moreover, in some cases, the **before-overlap** class was mistakenly recognized as **before** or **overlap** which degraded the overall recognition performance.

In the CR task, our second run (V2) performed best overall, indicating that a combination of machine learning and rule-based approaches is useful for this task. The main limitation of our approach is to use exhaustive (blind) search to extract possible pair relations. This results in many false positives and decreases the overall performance. Also, Moonstone rules are still under development, and will be further analyzed to increase accuracy.

Our aim was to benchmark existing tools and methods on this corpus. Adaptations of rule-based systems such as pyConText and HeidelTime proved insufficient on their own for the event and TIMEX3 subtasks compared to machine-learning based approaches, but were useful as feature input. Simple lexical features and cTAKES outputs were useful for the SVM and CRF classification approaches on the different subtasks. The narrative container relation is a very challenging task, requiring further feature engineering and analysis. We plan to further investigate and develop solutions where machine learning and rule-based approaches are combined, and to evaluate performance on other similar corpora.

Acknowledgments

The authors wish to thank the Mayo clinic and the 2015 Clinical TempEval challenge organizers for providing access to the clinical corpus. This work was partially funded by Swedish Research Council (350-2012-6658) and NLM R01 LM010964.

References

Steven Bethard, Philip Ogren, and Lee Becker. 2014. ClearTK 2.0: Design Patterns for Machine Learning in UIMA. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3289–3293, Reykjavik, Iceland, May.

Steven Bethard, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2015. Semeval-2015 task 6: Clinical tempeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

Brian E Chapman, Sean Lee, Hyunseok P Kang, and Wendy W Chapman. 2011. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. *J Biomed Inform*, 44(5):728–737.

Lee Christensen and Wendy W. Chapman. 2015. Moonstone. Manuscript in preparation.

Danielle Mowery, Sumithra Velupillai, Brett R. South, Lee Christensen, David Martinez, Liadh Kelly, Lorraine Goeuriot, Noemie Elhadad, Sameer Pradhad, Guergana Savova, and Wendy W. Chapman. 2014. Task 2: ShARe/CLEF eHealth Evaluation Lab 2014. In *CEUR Workshop Proceedings on CLEF 2014*, volume 1180, pages 31–42.

Sameer Pradhan, Noemie Elhadad, Brett R South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy W Chapman, and Guergana Savova. 2015. Evaluating the State of the Art in Disorder Recognition and Normalization of the Clinical Narrative. *J Am Med Inform Assoc.*, 22:143–154.

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)*, pages 394–397, Valletta, Malta, May.

Guergana K Savova, James J Masanz, Phillip 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.

Jannik Strötgen and Michael Gertz. 2013. Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, 47(2):269–298.

William IV 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 Özlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *JAMIA*, 20(5):806–813.

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.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 Task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 57–62, Stroudsburg, PA, USA.