Hitachi at SemEval-2016 Task 12: A Hybrid Approach for Temporal Information Extraction from Clinical Notes

Sarath P R[1], **Manikandan R**[1], **Yoshiki Niwa**[2]
[1]Research & Development Center, Hitachi India Pvt Ltd, India
[2]Hitachi Ltd, Center for Exploratory Research, Japan
{sarath,manikandan}@hitachi.co.in
yoshiki.niwa.tx@hitachi.com

Abstract

This paper describes the system developed for the task of temporal information extraction from clinical narratives in the context of 2016 Clinical TempEval challenge. Clinical TempEval 2016 addressed the problem of temporal reasoning in clinical domain by providing annotated clinical notes and pathology reports similar to Clinical TempEval challenge 2015. The Clinical TempEval challenge consisted of six subtasks. Hitachi team participated in two time expression based subtasks: time expression span detection (TS) and time expression attribute identification (TA) for which we developed hybrid of rule-based and machine learning based methods using Stanford TokensRegex framework and Stanford Named Entity Recognizer and evaluated it on the THYME corpus. Our hybrid system achieved a maximum F-score of 0.73 for identification of time spans (TS) and 0.71 for identification of time attributes (TA).

1 Introduction

Temporal information extraction has been a trending topic of research interest in the field of information extraction. It is crucial for improvement of systems used in number of applications ranging from question answering, search engines, text classification etc. to the systems that establish timelines and explicitly ground events that occurs in clinical narratives.

This work focuses on the automatic identification of time expressions from clinical texts. Time expressions are words and phrases that correspond to points or spans on a timeline, such as Dates, Time, Durations, Quantifiers, Set, and PrepostExp. Table 1 shows the time expression classes used in this work, with examples given for each class.

Time Class	Example
Date	February 2 2010, Friday morning
Time	5:30 PM, 20 minutes ago
Duration	For the next 24 hours, nearly 2 weeks
Quantifier	twice, three times
Prepost	postoperatively, post-surgery
Set	twice daily, weekly

Table 1: Examples of time expressions.

Research work on temporal information extraction has been carried out in both general NLP domain (Verhagen et al., 2007; Verhagen et al., 2010; UzZaman et al., 2013) and in clinical domain (Raghavan et al., 2012; Sun et al., 2013; Miller et al.,2013; Bethard et al., 2015; Velupillai et al 2015). In clinical domain, temporal information extraction has seen much research interest with assignment of medical events to time bins (Raghavan et al, 2012), i2b2 (Informatics for Integrating Biology and the Bedside) shared task (Sun et al., 2013) with the best performing system (Xu et al., 2013) using CRF (Conditional Random Field) and SVM (Support Vector Machines) and development of a hybrid system (Kovacevic et al., 2013) that used CRF and rule-based methods. Further tasks similar to i2b2 that involve temporal reasoning were also attempted in ShARe CLEF 2013/2014 eHealth challenges (Pradhan et al., 2015) and Clinical TempEval 2015 (Bethard et al., 2015; Velupillai et al., 2015).

Clinical TempEval 2016 continues with the task of temporal information extraction from clinical notes in similar lines to previously described

Proceedings of SemEval-2016, pages 1231–1236,
San Diego, California, June 16-17, 2016.

works. The 2016 Clinical TempEval challenge consisted of six subtasks (i) identification of spans of the time expression (TS) (ii) identification of spans of the event expression (ES) (iii) identification of attributes of time expression (TA) (iv) identification of attributes of event expression (EA) (v) identification of relation between an event and document creation time (DR) and (vi) identification of narrative container relation (CR). Participants were provided with THYME corpus (Styler et al., 2014) which consisted of de-identified clinical notes and pathology reports of cancer patients from Mayo clinic and corresponding annotation files. The participating systems were compared against two rule-based systems namely Memorize and Closest (only for CR task) which were used as baseline. The evaluation was done in two phases. In phase-1, only raw clinical text documents were given and participants were asked to identify time expressions, event expressions and temporal relations. In phase-2, access to raw text documents as well as manual event and time annotations was given and participants were asked to identify only temporal relations.

The 2015 Clinical TempEval results and discussion reveals that top results were mostly dominated by machine learning based systems (Bethard et al., 2015). However historical works on temporal extraction have achieved success using rule-based systems (Tang et al., 2012; Kovacevic et al., 2013). Thus in this work we developed a hybrid system leveraging the best out of machine learning and rule-based systems and evaluated this hybrid system against the THYME corpus.

The rest of the paper is organized as follows. In section 2 we describe the clinical TempEval 2016 data set and our approach to identify time expression. Section 3 and 4 explains results, describing the limitations of the methods used and issues that were present in the given THYME corpus.

2 Data and Method

The THYME corpus released for TempEval 2016 totaled 600 records which were divided into three sets namely train dataset, dev dataset and test dataset with 297, 150 and 153 records in each set respectively.

We tackled the tasks as named entity recognition (NER) problem with the aim to identify relevant text spans and assign classes to texts corresponding to the identified spans for which we developed a hybrid system that combines a rule-based method and a machine learning based method that relies on simple lexical, syntactic features and domain specific words. The rule-based system was developed using the Stanford TokensRegex framework (Chang et al., 2014) and the machine learning based system employed Stanford CRF classifier (Finkel et al., 2005), which are both available as part of the Stanford CoreNLP (Manning et al., 2014) tool set.

2.1 Rule-based approach

The rules for the rule-based approach were designed using combination of dictionaries and lexical formation formats of sentences. We started development of the rule-based system by evaluating the THYME corpus against existing Stanford SUTime (Chang et al., 2012) rule based tagger which is built on top of Stanford TokensRegex framework to understand its performance and shortcomings on THYME corpus. The obtained result showed that SUTime gave good result for Date when compared to other temporal classes. Hence we retained some of the rules for identification of Date from SUTime, while the rules for Time, Duration, and Quantifier and Set classes were developed as explained.

- **Time rules**: For time rules we manually crafted dictionaries that consisted of constituents of time expressions that were present as part of the given Time class annotations namely the word “time”, times of a day, time period qualifiers like “am”, ”pm” etc. and common time expression references (today, yesterday, previously, earlier, later, future etc.). Most of the annotations for Time class were phrases that contained word “time” or a time expression references prefixed by words such as “this”, “that” etc. and suffixed with time period qualifiers**.** All in all 16 different rules were developed to identify the time expressions using above said dictionaries.
- **Duration rules**: The duration rules used manually crafted dictionaries that consist of different constituents that were present as part of the annotations of duration class, namely weekdays and months, times of a

day and common time expression references (today, yesterday, previously, earlier, later, future etc.). In addition we created a dictionary of words that refers to time periods like (hours, minutes etc.) and their equivalent short forms and words that signify time references such as (since, dating back etc.). Totally we designed 43 different rules for identifying duration expressions using above said dictionaries.

- **Quantifier rules**: Most of the quantifier terms had an explicit dependency on words pertaining to domain specific or general events within in the same sentence where they are present .For example, in sentence ”Four cups of coffee in the morning” the term “Four” is the quantifier that quantifies the event “cups of coffee”. Similar to previous cases we handcrafted dictionaries that consist of domain specific and general events. Additionally a dictionary of commonly occurring suffixes such as “pack-year”, “pack-a-day” etc. that is part of the word that contained the quantifier value was created. Totally we created 15 unique rules for the quantifiers using above said dictionaries.
- **Prepost rules**: Prepost expressions had a common characteristics where each expressions begin with words such as “pre”, ”post”, ”intra”, ”prior” etc. followed by a domain specific event term such as “operative” etc. Further we observed very few variants of prepost expressions being present in dev dataset. However to avoid unseen words that might follow the above mentioned prefixes in test dataset, we extracted domain specific words that commonly follow these prefixes by mining ICD9 website and in addition we also created a dictionary of various words that is related to surgical procedures by mining Wikipedia. Using above said dictionaries and prefix words we created 5 rules to extract prepost expressions. Once the quantifiers and prepost expressions were extracted we have set of post processing modules that were designed to improve the spans by removing certain words from extracted expressions.
- **Set rules**: The set rules used previously handcrafted dictionaries for time rules. In addition we created a dictionary containing words such as “annual”, ”weekly”, ”monthly”, “daily” etc. that qualifies a Set. Most of the set expressions were either single words from previously described dictionary or a simple sentence of form “XX-times-a-{time period qualifier}”,”XX-{ time period qualifier }” etc. where XX is numeric quantifier. Totally 20 different rules were designed for identifying set expressions.

The results of the rule-based system that was developed using above mentioned rules on test dataset are as shown in Table 2.

Subtask	P	R	F1
TIMEX3	0.415	0.629	0.500
TIMEX3: Span	0.433	0.655	0.522
TIMEX3: Class	0.415	0.629	0.500
TIMEX3: Date	0.457	0.690	0.550
TIMEX3: Duration	0.298	0.455	0.360
TIMEX3: Prepost	0.986	0.637	0.774
TIMEX3: Quantifier	0.256	0.348	0.295
TIMEX3: Set	0.536	0.541	0.538
TIMEX3: Time	0.110	0.378	0.171

Table 2: Results for TS and TA tasks on the test dataset using rule-based system.

2.2 Machine learning approach

We created separate models for all six types of time classes using Stanford CRF classifier which is an implementation of arbitrary order linear chain conditional random field classifier. The data for training consisted of 447 records (train and dev datasets) which were preprocessed by tokenizing using the Penn Tree bank tokenizer and BIO encoding each of the tokens.

We used simple features such as N-Grams, word shape features, word window of size ±1, sequence words from Stanford NER Feature factory for creating the model and tested using 153 records from test dataset. The results that were obtained are as shown in Table 3.

Subtask	P	R	F1
TIMEX3	0.79	0.655	0.720
TIMEX3: Span	0.821	0.669	0.737
TIMEX3: Class	0.798	0.655	0.720
TIMEX3: Date	0.823	0.749	0.784
TIMEX3: Duration	0.650	0.455	0.535
TIMEX3: Prepost	0.969	0.832	0.895

TIMEX3: Quantifier	0.500	0.167	0.250
TIMEX3: Set	0.732	0.369	0.491
TIMEX3: Time	0.385	0.167	0.233

Table 3: Results for TS and TA tasks on the test dataset using machine learning system (CRF).

3 Results

During system development phase we were able to see that rule based system performed well for Quantifier and Set while CRF performed well for the rest of the classes. Hence for the submission runs we used a hybrid of our rule-based and machine learning based systems (CRF) with the test data. For Quantifier and Set classes we used rule-based systems and for Date, PrepostExp, Duration, and Time class we used CRF classifier models.

Subtask	**P**	**R**	**F1**
TIMEX3	0.759	0.671	0.712
TIMEX3: Span	0.781	0.685	0.73
TIMEX3: Class	0.759	0.671	0.712
TIMEX3: Date	0.823	0.749	0.784
TIMEX3: Duration	0.65	0.455	0.535
TIMEX3: Prepost	0.969	0.832	0.895
TIMEX3: Quantifier	0.256	0.348	0.295
TIMEX3: Set	0.536	0.541	0.538
TIMEX3: Time	0.385	0.167	0.233

Table 4: Hitachi team results (run-1) for TS and TA tasks in Clinical TempEval 2016.

For run-1 we used 447 records (train & dev datasets) for training our hybrid system and the result which was obtained when evaluated on test data is shown in Table 4. For run-2 we decided to do an estimation of how well the CRF model has been trained and its property: accuracy dependency on number of training records. Hence we replaced the CRF classifier with models trained only on 297 records (train dataset). The results on test data for run-2 were similar to run-1 except a drop in overall F-score of 0.1.

4 Discussion

Our hybrid system outperformed the baseline systems for TS and TA tasks. We also obtained results that were above the median result of the challenge as shown in Table 5.

Subtask	**P**	**R**	**F1**
TIMEX3: Span (Hitachi run-1)	0.781	0.685	0.730
TIMEX3: Span (Hitachi run-2)	0.781	0.668	0.720
TIMEX3: Span (TempEval 16 baseline)	0.744	0.428	0.551
TIMEX3: Span (TempEval 16 top score)	0.84	0.75	0.795
TIMEX3: Span (TempEval 16 median score)	0.779	0.539	0.637
TIMEX3: Class (Hitachi run-1)	0.759	0.671	0.712
TIMEX3: Class (Hitachi run-2)	0.758	0.654	0.702
TIMEX3: Class (TempEval 16 baseline)	0.746	0.413	0.532
TIMEX3: Class (TempEval 16 top score	0.815	0.735	0.772
TIMEX3: Class (TempEval 16 median score)	0.755	0.499	0.618

Table 5: Comparison of Clinical TempEval 2016 results.

Table 2 shows the results of our rule-based system on test data for which we observe an F-score of 0.50 with date, prepost and set expression having recall higher than 0.5. During the system development we tuned the rule-based components towards the patterns of temporal expressions that were pre-identified in the training and dev dataset, but there were words such as “time” for Time class and “MO”, “hrs” etc. for Duration class which led to increase in the number false positives thereby reducing the precision and overall F-score. Furthermore, the rule-based method, extracted many time expressions like “7:45 AM”,” 24-May-2010 15:12:00”, “10 Units” etc. that were valid but not present in the annotations which led again to reduction in F-scores.

Table 3 shows the results of our machine learning system (CRF) on test data. The result obtained using CRF classifier is consistent with the previous works that is based on CRF. CRF gave an average F score of 0.72 on test data. In fact CRF gave very good results for all classes except those belonging to Time and Quantifier class. For instance when tested with the test data for Duration class, CRF predicted 140 patterns out of which 91 were correct leading to a precision of 0.65. For the same attribute rule-based system extracted

305 patterns out of which only 91 were correct giving a low precision of 0.298.

We started our system development to understand how rule-based method and CRF performs individually for extraction of time expression present in the dev dataset of the THYME corpus. We were able to observe that the rule-based and machine learning systems gave good results only on subset of time expressions when used individually which can also be observed with the result on test dataset shown in Table 2 and 3. Further, based on results from Table 3 and 4 we can say that CRF alone has higher performance than the hybrid system, however we observed opposite results during system development process using dev dataset. Thus for the final submission we developed a hybrid system of rule-based and CRF by combining top performing systems on dev data, which lead to the results shown in Table 5.

Adaptation of Stanford TokensRegex framework for rule-based system performed fairly well giving an average F-score 0.5 for test data. Yet our rule-based method of the hybrid system had major limitation where our rules were highly syntax dependent which was unavoidable. Simple lexical features were useful for CRF classification approaches on TS and TA tasks. Further, the overall impact of reduction in training data for run-2 was negligible which is evident from the result. Also we observed that performance of CRF for classes like Quantifier remained unimproved even with addition of higher level syntactic features, which is evident from results in Table 3. Thus our aim of evaluating our hybrid system against THYME corpus which was developed using Stanford TokensRegex Framework and Stanford CRF Classifier was successful. As a future work we plan to evaluate performance of our hybrid system on other similar corpora and explore various strategies to combine rule-based and machine learning methods.

Acknowledgements

We thank Mayo clinic and clinical TempEval organizers for providing access to THYME corpus and other helps provided for our participation in the competition.

References

Aleksandar Kova˘cevi´c, Azad Dehghan, Michele Filannino, John A Keane, and Goran Nenadic. 2013. Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. *Journal of the American Medical Informatics Association, 20(5):859– 866.*

Angel X. Chang, and Christopher Manning. 2012. SUTime: A library for recognizing and normalizing time expressions. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).*

Chang, Angel X., and Christopher D. Manning. TokensRegex: Defining cascaded regular expressions over tokens. 2014. *Technical Report CSTR 2014-02, Department of Computer Science, Stanford University.*

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370.*

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit.In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.*

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 Fourth International Workshop on Semantic Evaluations (SemEval- 2007), pages 75–80, Prague, Czech Republic, 90 June. 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, Uppsala, Sweden, July. Association for Computational Linguistics.*

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. ACL*

Preethi Raghavan, Eric Fosler-Lussier, and Albert M Lai. 2012. Temporal classification of medical events. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, pages 29–37. Association for Computational Linguistics.*

Steven Bethard, Leon Derczynski, Guergana Savova, and James Pustejovsky. 2015. SemEval-2015 Task 6: Clinical TempEval. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 806-814*.

Sumithra Velupillai, Danielle L Mowery, Samir Abdelrahman, Lee Christensen, and Wendy W Chapman. 2015. BluLab: Temporal Information Extraction for the 2015 Clinical TempEval Challenge. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages-815-819*.

Tang B, Wu Y, Jiang M, Chen Y, Denny JC, Xu H. A hybrid system for temporal information extraction from clinical text. *Journal of the American Medical Informatics Association : JAMIA*. 2013;20(5):828-835. doi:10.1136/amiajnl-2013-001635.

Timothy Miller, Steven Bethard, Dmitriy Dligach, Sameer Pradhan, Chen Lin, and Guergana Savova. 2013. Discovering temporal narrative containers in clinical text. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing, pages 18–26, Sofia, Bulgaria, August. Association for Computational Linguistics.*

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.*

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.*

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.*