XJNLP at SemEval-2017 Task 12: Clinical temporal information extraction with a Hybrid Model

Yu Long
Xi’an Jiaotong University
yulongxbdx@163.com

Zhijing Li
Xi’an Jiaotong University
tokyojackson@stu.xjtu.edu.cn

Xuan Wang
Xi’an Jiaotong University
m18829236255@163.com

Chen Li
Xi’an Jiaotong University
cli@xjtu.edu.cn

Abstract

Temporality is crucial in understanding the course of clinical events from a patient’s electronic health records and temporal processing is becoming more and more important for improving access to content. SemEval 2017 Task 12 (Clinical TempEval) addressed this challenge using the THYME corpus, a corpus of clinical narratives annotated with a schema based on TimeML2 guidelines. We developed and evaluated approaches for: extraction of temporal expressions (TIMEX3) and EVENTs; EVENT attributes; document-time relations. Our approach is a hybrid model which is based on rule based methods, semi-supervised learning, and semantic features with addition of manually crafted rules.

1 Introduction

Extraction and interpretation of temporal information from clinical text is essential for clinical practitioners and researchers. 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. SemEval-2017 Task 12 (Clinical TempEval) is a direct successor to 2016 Clinical TempEval. Clinical TempEval is designed to ad-dress the challenge of understanding clinical timeline in medical narratives and it is based on the THYME corpus which includes temporal an-notations.

Researchers have explored ways to extract temporal information from clinical text. Lee et al. (2016) developed an approach based on linear and structural (HMM) support vector machines using lexical, morphological, syntactic, discourse, and word representation features. P R, Sarath et al. (2016) used a hybrid approach(rule-based and machine learning) for temporal information extraction from clinical notes. Velupillai et al. (2015) developed a pipeline based on ClearTK and SVM with lexical features to extract TIMEX3 and EVENT mentions. Most of the participants of these challenges used CRF and SVM for event and time expression extraction with features including the information gathered from different resources like UMLS (Unified Medical Language System), output of TARSQI toolkit, Brown Clustering, Wikipedia and Metamap (Aronson and Lang, 2010). Those machine-learning methods are complex and they cost much time to run. However, they can be not only flexible but also convenient when compared to the handcrafting label. Others also used some rule based methods, which are fast but not flexible enough. It seems that the combination of those two methods may gain the better result. Since 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), Galescu and Blaylock (2012) and so on. Recently, interest in temporal processing has moved forward in two directions: cross-document timeline extraction (Minard et al., 2015) and domain adaptation (Sun et al., 2013; Bethard et al., 2015). Based on the analysis above, our hybrid model utilize machine learning techniques and crafted rules which contains SVM (Support Vector Machine) classifier and RNN (Recurrent Neural Networks) classifier to extract Temporal Information from Clinical documents and make classifications.

2 Data and Method

2.1 Data

We use THYME corpus for training and evaluating the methods, which consists of clinical and pathology notes of patients with colon cancer and brain cancer from Mayo Clinic. The THYME corpus is split into training, development, and test sets based on patient number, with 50% in training and 25% each in development and test sets. Table 2 shows the distributions of the different time and event classes in the THYME corpus. The training data about colon cancer contains 3,833 time expressions and 38,890 events, the development data contains 2,078 time expressions and 20,974 events. The training data about brain cancer contains 350 time expressions and 2,557 events.

Table 1: different time and event attributes in the

	attribute	Coloncancer-Train	Brain-cancer-Train	Coloncancer-Dev
EVENT	Documents	293	30	147
	ASPECTUAL	546	51	246
	EVIDENTIAL	2,206	85	1,314
	N/A	36,185	2,421	19,414
	MOST	96	2	45
	LITTLE	143	18	65
	N/A	38,698	2,537	20,864
	POSITIVE	34,832	2,386	18,795
	NEGATIVE	4,105	171	2,179
	ACTUAL	35,781	2,172	22,647
	HEDGED	889	81	443
	HYPOTHETICAL	1,656	88	829
	GENERIC	611	216	611
TIMEX	Date	2,588	204	1,422
	Duration	434	29	200
	PrePostExp	313	37	172
	Set	218	13	116
	Quantifier	162	9	109
	Time	118	58	59

THYME3 corpus

The data of colon cancer are more than others and the training data of brain cancer is too little but the test data is all about brain cancer, so the task will focus on domain adaptation. We can also see the unbalanced data distribution, for example, the data of N/A is 38,698, but the data of MOST is only 96, and maybe unbalanced data will have an impact on the results. We used the development set for optimizing learning parameters, then combined it with the training set to build the system used for reporting results in Section 4.

2.2 Task Description

Clinical TempEval 2017 was focused on designing approaches for information extraction in the clinical domain.There were 6 different tasks which are listed in Table 2.

Clinical TempEval is designed to address the challenge of understanding clinical timeline in medical narratives and it is based on the THYME corpus which includes temporal annotations.

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 2: Tasks of clinical TempEval 2017

For extracting temporal information from clinical text, we utilize semi-supervised learning algorithms (SVM and RNN) 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.

3 Methodology

Our approach to the tasks is a hybrid model that is based on rule based methods and 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:

3.1 TIMEX3 Span Detection and Time Expression Attribute Identification

Our tasks are about time expression span detection (TS) and time expression attribute identification (TA), which means that we should first extract the time expression and then identify which class it belongs to. As for time span, we use the rule based methods to detect the boundary of the time expression. We use Stanford NLP package to do the preprocessing and we normalize the digital expressions after it, we change every character to “0” as long as it is digit. (e. g. we normalize the "12:13" to "00:00".)

For the rule based methods, firstly we find all the prepositions, according to our experience and experimental statistics, we extract five tokens behind their own prepositions. Since we thought that many time expressions always show up behind a preposition, we then judge whether those five words are related to time expressions. We define a time dictionary to list the words which we think can be a part of the time expressions, like "month", "week, "day", "hour", "May", "Monday", "morning", "once" and so on. Next, we contrast the five tokens with time dictionary, and find whether it can represent a date or a precise time. Finally, we extract all the continuous tokens that we thought may relate to the time expressions (if there is a definite article before those tokens, extract it as well). There exist some expressions do not after a preposition and only contain one word and most of them have the same prefix like "pre", "post", "peri". So we use this prefix rule to find the remain expressions.

We also use the rule based methods to identify the classes of the time expression. And here are some examples of the rules for each class:

Class	Rules
Date	1999-11-08, yesterday, last Saturday, in 3 years, 3 months ago...
Duration	for 3 days, July to August, since last summer....
PrePostExp	post, preoperative, prior to the surgery....
Set	Twice per day, 3 times a day...
Quantifier	Twice, once...
Time	13:56, in the morning....

Table 3: examples of rules for each time expression class

3.2 Event Extraction Task

In this task, we need to extract medical events from the clinical text and identify attributes of the events which are showed in table 1.

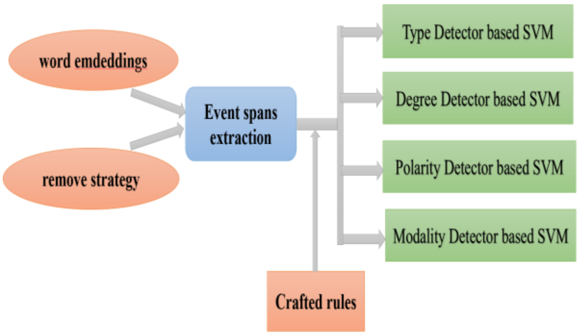

Figure 1 Event Extraction Architecture

Figure 1 illustrates the architecture of our EVENTs extraction system. First, we create word embeddings using the Wikipedia database. Then we extract event spans with a SVM classifier and a remove strategy. Finally we detect type, degree, modality, and polarity using four separate SVM classifiers and crafted rules.

3.2.1 Event Spans (ES) Extraction

To extract EVENT spans, first we train a separate Support Vector Machine to complete prediction. Then we make a colon corpus about colon cancer which comes from training data and Wikipedia. Finally, we remove the events which exist in the colon corpus from the prediction result.

The major feature we used for training the SVM classifier is word embeddings. We trained all word embeddings in this document with word2vec (Mikolov et al., 2013) using the Skipgram model on a text window size of 2 tokens, to obtain words vector representations of dimension 50. We also try to use the words vector representations of dimension 300, but the result is unexpected.

3.2.2 Identifying EVENTs Attributes (EA)

Table 1 shows the EVENTs attributes. Assigning these attributes to one of its values is an classification task. We train four separate Support Vector Machines for each attribute to classify their respective classes. We also use word embeddings as the major features for training separate SVM classifier for each attribute.

Furthermore, according to our observations of the corpus, different types of event mentions may show different rules. For instance, events with EVIDENTIAL type are usually represented with verbs such as “showed”, “reported”, “found”, in contrast, the events with N/A type that are usually represented medical terms such as “nausea”, “chemotherapy” or “colonoscopy”. So we create such rules to help classifications.

3.3 Document-time Relation (DR)

Document-time relations (DR) are specific attributes of EVENTs indicating their temporal relation with the document creation time. There are 3 different types of DRs, namely, BEFORE, AFTER, and OVERLAP. For identifying the DR attribute types, we use RNN. RNN makes up for the inaccuracy of the convolution kernel and the pool size in the process of text processing, therefore, the generated RNN classifier has higher accuracy for text classification. We train classifier for each DR type using an set of features to what was used for EVENTs attributes detection. Verb tense and the modals in the sentence are also indicative of the sentence tense and can help in identifying the document-time relation. Figure 1 describes the additional features that we use for DR extraction. In addition to the base features, we consider features specific to the EVENTs annotation. We furthermore expanded the features by considering contextual features from the sentence and nearby time and date mentions. We try to optimized the RNN classifier--thread level speculation. Replace the calculated results of the other core to be weighted with speculative value, in that way, the parallel computing can be carried out smoothly. We used this method to classify the colon cancer data with golden annotations, the results are shown in the following table.

DR	P	R	F1
RNN	0.69	0.71	0.70
RNN+	0.90	0.91	0.90

Table 4: Document-time Relation of cancer data

From this table, we can see the value of precision, the value of recall and the value of F1 are relatively high, so the Optimized RNN classifier is effective. But we do not know whether it is suitable for the brain cancer data.

4 Experiments and Results

The 2017 Clinical TempEval task consisted of two evaluation phases. Phase1 is unsupervised domain adaptation and phase 2 is supervised domain adaptation. In phase 2, we participated in all tasks, except for CR.

We report the results on the test set for all subtasks, Results have been computed in terms of Precision (P), Recall (R) and F1. For comparison we will also report the maximum scores of the participating systems.

Subtask	P	R	F1
TIMEX3_SPAN	0.33	0.52	0.41
TIMEX3_Class	0.29	0.45	0.35

Table 5: results for TS and TA subtasks

However, the result is less than satisfactory. Table 5 shows the final result. We compared our results with the best results on the Semeval website. (https://competitions.codalab.org/) We think there are three reasons: First, our methods always extract two different expressions as one if they are very close to each other. Secondly, our dictionary is too small to cover enough words. Thirdly, we extract most of words in the raw text that have the prefix "pre", "post", "peri", but some of them are not time expressions. As for TA, we think that we only focus on the time expression itself but ignore much semantic information.

Subtask	P	R	F1
ES	0.55	0.69	0.61
Type	0.53	0.66	0.59
Degree	0.54	0.67	0.56
Polarity	0.49	0.61	0.54
Modality	0.46	0.57	0.51

Table 6: results for EVENTs subtasks

The results for EVENTs subtasks also show lower performance in comparison with the result of best system. Error analysis are as bellowed:

Firstly, we don't use a good and effective domain adaption method, and we do not have an effective way to solve the unbalanced data. Secondly, we don't integrate more domain specific features. Thirdly, in the process of Events Attributes identification, we ignore the importance of context analysis and Sentiment analysis. For example, "bleeding" can be the positive class of the Polarity attribute, and it also can be the negative class. This is up to the context analysis. In addition, we create word embeddings using the Wikipedia database. The temporal information from clinical is professional. So we need to use more database about clinic to improve the performance of the word embeddings. In the future, we plan to further improve our system to show higher performance based on the observations above.

Subtask	P	R	F1
DR	0.29	0.36	0.32

Table 7: results for DR subtasks

We use the results of EVENT extraction to forecast the document-time relation of brain cancer. So the results of EVENT_span and TIMEX3_span are very important, and we do not add the domain adaptation, so the result of DR of brain cancer are relatively low, the detailed results are shown in table7. We have identified some errors: first, wrong output of the pre-processing modules, especially the parsing process. Second, limitations of the features selected. Third, lack of domain specific knowledge.

5 Discussion and Conclusions

SemEval 2017 task 12 (Clinical TempEval) was focused on temporal information extraction from clinical narratives. Our methods employed rule based methods and machine learning classification scheme for all the tasks except for CR based on various sets of syntactic, lexical, and semantic features. We illustrated that incorporating manually crafted extraction rules improves results, but the rules should be improved.

For TIMEX3 subtasks, our approach was clearly not the best solution as our rules are simple and not perfect so that the system cannot obtain the high score. For EVENTs subtasks, our system is not ideal for unbalanced data classification, and we will enhance its effectiveness. For DR subtask, we showed that the optimized classifier can improve the accuracy but we do not know whether it is suitable for the brain cancer data. Besides, we do not consider the domain adaptation and our features were minimal. There are many options to improve the system, ranging from fine tuning the pre-processing phase in order to avoid offset misalignments, to the generation of better features for the ES and DR subtasks. In future work, we aim to implement all the improvements mentioned above.

Acknowledgement

This work has been supported by "The Fundamental Theory and Applications of Big Data with Knowledge Engineering" under the National Key Research and Development Program of China with grant number 2016YFB1000903; Project of China Knowledge Centre for Engineering Science and Technology; Innovation team of Ministry of Education (IRT_17R86); Project of China Mobile Research (MCM20160404); Project of thousands of young people plan.

References

Rodriguez P, Wiles J, Elman J L. A RNN that learns to count[J]. Connection Science, 1999.

Wang W J, Liao Y F, Chen S H. RNN-based prosodic modeling for mandarin speech and its application to speech-to-text conversion[J]. Speech Communication, 2002, 36(3-4):247-265.

Cho K, Merrienboer B V, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[J]. Computer Science, 2014.

Jain A, Zamir A R, Savarese S, et al. Structural-RNN: Deep Learning on Spatio-Temporal Graphs[J]. Computer Science, 2015.

Namikawa J, Tani J. A model for learning to segment temporal sequences, utilizing a mixture of RNN experts together with adaptive variance[J]. Neural Networks, 2008, 21(10):1466-1475.

Shao H, Nonami K, Wojtara T, et al. Neuro-fuzzy position control of demining tele-operation system based on RNN modeling[J]. Robotics and Computer-Integrated Manufacturing, 2006, 22(1):25-32.

Routraya G, Kanungo P. Genetic Algorithm Based RNN Structure for Rayleigh Fading MIMO Channel Estimation[J]. Procedia Engineering, 2012, 30(30):77-84.

Chung J, Jang H, Jung K H, et al. Parallel approach for processing itinerary-based RNN queries in object tracking WSNs[J]. Telecommunication Systems, 2014, 55(1):55-69.

Patnaik P R. A RNN for a fed-batch fermentation with recombinant Escheria coli subject to inflow disturbances[J]. Process Biochemistry, 1997, 32(5):391-400.

Boronat M, Corma A, Gonz álezarellano C, et al. Synthesis of Electron-Rich CNN-Pincer Complexes, with N-Heterocyclic Carbene and (S)-Proline Moieties and Application to Asymmetric Hydrogenation[J]. Organometallics, 2014, 29(1):159-172.

Weiyi Sun, Anna Rumshisky, and OzlemUzuner. 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.

Alan R Aronson and Franc, ois-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, James Pustejovsky, and Marc Verhagen. 2015. Semeval-2015 task 6: Clinical tempeval. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Association for Computational Linguistics.