ULISBOA at SemEval-2017 Task 12: Extraction and classification of temporal expressions and events

Andre Lamurias[1], Diana Sousa[1], Sofia Pereira[1], Luka A. Clarke[2] and Francisco M. Couto[1]

[1]LaSIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
[2]BioISI: Biosystems & Integrative Sciences Institute, Faculdade de Ciências, Universidade de Lisboa, Portugal

Abstract

This paper presents our approach to participate in the SemEval 2017 Task 12: Clinical TempEval challenge, specifically in the event and time expressions span and attribute identification subtasks (ES, EA, TS, TA). Our approach consisted in training Conditional Random Fields (CRF) classifiers using the provided annotations, and in creating manually curated rules to classify the attributes of each event and time expression. We used a set of common features for the event and time CRF classifiers, and a set of features specific to each type of entity, based on domain knowledge. Training only on the source domain data, our best F-scores were 0.683 and 0.485 for event and time span identification subtasks. When adding target domain annotations to the training data, the best F-scores obtained were 0.729 and 0.554, for the same subtasks. We obtained the second highest F-score of the challenge on the event polarity subtask (0.708). The source code of our system, Clinical Timeline Annotation (CiTA), is available at https://github.com/lasigeBioTM/CiTA.

1 Introduction

This paper presents our system CiTA (Clinical Timeline Annotation) to participate in the SemEval 2017 Task 12: Clinical TempEval challenge. Our team participated in the subtasks corresponding to the identification of the following properties: time expression spans, event expression spans, time expression attributes, event expression attributes. Time expressions had only one attribute, type, which could be DATE, TIME, DURATION, QUANTIFIER, PREPOSTEXP or SET. Event attribute identification consisted of four attributes: type (N/A, ASPECTUAL or EVIDENTIAL), polarity (POS or NEG), degree (N/A, most or little) and modality (ACTUAL, HEDGED, HYPOTHETICAL or GENERIC).

For this challenge, we developed a system, named Clinical Timeline Annotation CiTA[1], which uses IBEnt (Lamurias et al., 2015) to identify the text spans of time and event entities based on machine learning and semantic techniques. CiTA also incorporates hand-crafted rules to assign the attributes to each entity. We trained one classifier for each entity type using Conditional Random Fields (CRF) and developed a set of rules for each attribute, based on the training data available at each phase. This paper describes the features and resources used for each subtask, presents our results and discusses the main issues found. CiTA is publicly available in a GitHub repository [2].

2 Methods

A corpus of clinical, pathology and radiology notes from the Mayo Clinic was available to the participants. This corpus contained notes for the source domain (colon cancer) and for the target domain (brain cancer). Each document was manually annotated with time and event expressions, as well as their attributes. The annotators and adjudicators followed a set of guidelines which were also available to the participants. During Phase 1 only annotated colon cancer reports were avail-

ălamurias@lasige.di.fc.ul.pt

[1]http://labs.fc.ul.pt/cita/
[2]https://github.com/lasigeBioTM/CiTA

able, and in Phase 2 thirty annotated brain cancer documents were also released.

The colon cancer dataset was partitioned in 3 sets: train, development and test. We trained the system with the train and development set, and optimized with the test set. In Phase 2, we enhanced the classifiers by adding the brain cancer annotated documents. We ignored sections of the colon cancer documents that were not annotated due to the guidelines.

2.1 Event / Time Entity Span Identification

For both ES and TS subtasks we trained CRF classifiers on the training data annotations. We trained a CRF classifier for events and another for time expressions, using CRFSuite (Okazaki, 2007). These classifiers identified only the spans of the entities so that we can evaluate and improve the results of this subtask before classifying the entity attributes. This is justified in the context of the competition since the attribute classification subtasks are dependent on the span identification subtasks, and a poor performance on the span identification subtasks would affect the other subtasks.

We used a set of common features for time and event expressions, based on previous experiments, that explored linguistic, orthographic, morphological and contextual properties of the tokens (Table 1). For most features, we considered a contextual window of size one, i.e., the value of the same feature for the previous and next token. Lemma and Part-of-Speech tags were obtained using Stanford CoreNLP (Manning et al., 2014).

Furthermore, we selected specific features for time and event expressions. For time expressions, we used the NER tags given by SUTime (Chang and Manning, 2012), part of Stanford CoreNLP. SUTime is able to detect general time and date expressions, which is the case of some of the time expressions in the gold standard. For the event classifier, we matched each word in the gold standard to a Unified Medical Language System (UMLS) concept and used it as a feature if the confidence level was higher than 0.8. The matching was performed using LDPMAP (Ren et al., 2014). Many words were matched to UMLS since it is large vocabulary. However, by applying a high threshold, we ensure that only high quality matches are considered.

During Phase 2, we analyzed some errors made by the colon cancer classifiers on the brain cancer training set. To overcome these errors, we automatically created a list of common false positives and false negatives for time and event expressions. We applied the false positives list to the output of the CRF classifiers as a filter, and performed a dictionary search with the false negatives list to identify missed entities. We used these lists only on Run 2 of our Phase 2 submission.

Feature	Context window	Entity
Prefix sizes 2-4	-1/1	All
Suffix sizes 2-4	-1/1	All
Contains number	0	All
Case	-1/1	All
Lemma	-1/1	All
POS tag	-1/1	All
Word class	-1/1	All
SUTIME tag	-1/1	Time
POS tag	-2/2	Event
UMLS	-1/1	Event

Table 1: Features used for TS and ES subtasks.

2.2 Event / Time Entity Attribute Classification

Each event and time entity identified by CiTA was then classified according to the attributes defined by the task. To this end, we established a set of rules for each attribute using regular expressions. These rules were developed according to the annotation guidelines and training data. The rules developed for modality and polarity attributes were based on the context windows of each event. Furthermore, we chose the default values of each attribute based on the frequency of each value on the brain cancer annotations.

We found that several expressions used in the context window of the event affected its modality and polarity. For polarity, *avoid*, *absent* and *not* indicated a negative polarity. If the context of the event did not include any of the expressions of our list, we classified it as positive (95.9% of the cases). For modality, we selected *ACTUAL* as the default value (84.9% of the cases), since it is the most frequent value.

To choose the size of the context windows, various sizes were tested, both to the left and right of the event. We noticed that if we extended the window too much, some expressions that did not affect the event would be matched. However, shorter

context windows would not include the relevant expressions. We fixed the window size of 5 for both polarity and modality. If the conjunctions *but* or *with* were found in the context window, we cut the window at that point. These conjunctions change the subject of the sentence from the respective event, and all words afterwards were ignored. Furthermore, we ignored any expressions that affected the polarity of an event if there was another event between the expression and that event. For example, if the expression *not* appears in the left context window of event A, but event B also appeared in the same window, between *not* and event A, then event A was classified as positive.

For the other attributes (event type and time type) we chose a different approach. Although we tried to formulate rules based on the context windows of each event and time entity we realized that it was more efficient to make a direct match between the attribute and the event or time entity. To classify type of events, as it was said on the set of the guidelines available to the participants, we realized that specific groups of verbs indicated a certain modality, for example, *evidence* (EVIDENTIAL) or *starting* (ASPECTUAL), making it easy to recognize which verbs belong to this class. We developed rules based on each modality class, except for the default value (N/A) (94.7% of the cases). The rules used to classify the type of time expressions were slightly different. We had to identify which of the six attributes was the default or the one that included the widest amplitude of expressions. We started by making rules for each attribute by identifying the patterns in the gold standard, quickly realizing that the default attribute was DATE (59.3% of the cases). So we focused our attention on the definition of the other five attributes (PREPOSEXP, SET, TIME, DURATION and QUANTIFIER) by matching the different type of time expressions and possible variations to each appropriated attribute.

3 Results

We submitted one run during Phase 1 and two runs during Phase 2. While during Phase 1 we only had access to source domain annotations, some target domain annotations were released for Phase 2. Hence, we were able to improve the performance of CiTA in relation to the target domain during Phase 2. Table 2 shows the official results for Phase 1 and Phase 2 Run 1 and 2. For each run, we present the precision, recall and F1-score obtained in each subtask.

Compared to the results of Phase 2, Phase 1 results were lower, particularly for Time span identification ($\Delta = 0.069$). The false positive filter applied on Phase 2 Run 2 improved the precision of the time span subtask, although at the expense of a lower recall. On the event span subtask it results in a lower precision, with almost no effect on recall. In both phases, the results for time expressions were lower than for events.

The results of the time and event attributes are shown in combination with the span identification. This means that an entity is considered positive if both the span and attribute are found in the gold standard. Hence, we can evaluate the effect of the rules on the test set by comparing the scores of each attribute to the span identification score. Furthermore, we evaluated the accuracy of the rules on the colon cancer and brain cancer train sets (Table 3). We assumed that the attribute value was correct if it matched the gold standard. Table 3 shows the results obtained using the rules developed for the second phase, which were tuned for the brain cancer data sets.

4 Discussion

The main challenge of this task was to adapt a system developed for a specific source domain to a different target domain. Systems trained on a specific domain, either using hand-crafted rules or machine learning, are biased for that domain. In real world scenarios, information extraction systems need to be able to perform well in multiple domains. Although at first it seemed like the only difference between the source and target domains was the type of cancer, we observed that the reports and annotations were also different in terms of form of the documents and terms used. These differences contributed to lower scores obtained on the target domain test set, when compared to the source domain test set used in the previous edition of this task (Bethard et al., 2016). Even using the brain cancer train set available during Phase 2, our best F1 score on the event span subtask was 0.16 lower than on the colon cancer test set.

Comparing to the other teams that submitted results to this task, our submission performed better on the event expressions subtasks. On Phase 1, we are the third best team on all event subtasks in terms of F1 score. On Phase 2, we are in second

	Phase 1			Phase 2 Run 1			Phase 2 Run 2			Phase 2 Top F1		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Event span	0.618	0.765	0.683	**0.649**	**0.831**	**0.729**	0.637	0.830	0.721	0.69	0.85	0.76
Event modality	0.548	0.679	0.607	**0.571**	**0.731**	**0.641**	0.561	0.730	0.634	0.63	0.78	0.69
Event degree	0.610	0.756	0.675	**0.642**	**0.821**	**0.720**	0.630	0.820	0.713	0.68	0.84	0.75
Event polarity	0.600	0.744	0.664	**0.630**	**0.807**	**0.708**	0.619	0.806	0.700	0.68	0.83	0.75
Event type	0.598	0.741	0.662	**0.629**	**0.805**	**0.706**	0.617	0.804	0.698	0.68	0.83	0.75
Time span	0.441	0.538	0.485	0.517	**0.598**	**0.554**	**0.520**	0.588	0.552	0.57	0.62	0.59
Time type	0.393	0.479	0.432	0.483	**0.559**	**0.518**	**0.485**	0.548	0.515	0.54	0.59	0.56

Table 2: Results of our submission and for each task the results from the top F1 score submission of Phase 2. Notice that *Time type* represents *Time span + Class* column of the results table published by the organizers.

Gold standard	Colon	Brain
Event modality	**0.914**	0.891
Event degree	**0.995**	0.993
Event polarity	**0.969**	0.968
Event type	**0.972**	0.962
Time type	0.890	**0.971**

Table 3: Accuracy obtained using the rules developed for each attribute, on the colon cancer and brain cancer train sets.

place on the Event polarity subtask, maintaining the third place on all the other event subtasks, except modality. In time span and type we are in third place in terms of recall.

4.1 Error Analysis

Some of the more persistent errors while classifying the type of time entities were that some of the time expressions presented in the gold standard had double attribution. For example, *time* sometimes appeared classified as TIME and in others as DURATION, and *daily* was classified as DATE, SET and QUANTIFIER. Although initially we tried to introduce context windows of each time expression to help us solving this systematic error, we realized that there was no explicit difference in the context of most of these words, so introducing context windows only harmed our efforts to achieve better results.

Some event attributes were incorrectly classified due to the developed context window rules. For example, in *not limited to loss of appetite*, *appetite* was incorrectly classified with negative polarity since it had *not* in its left context window, and no event in-between. In some cases, the rule we implemented to ignore negation expressions between events resulted in incorrect positive polarities. In the expression *no second lesion seen in the brain*, *no* did not affect *seen* because another event, *lesion* appeared in its context window. However *seen* was supposed to be classified as negative.

One limitation of a rule-based approach is that it is necessary to take into account every expression that might affect an attribute. Since we had a limited amount of target domain training data, we missed some cases where more complex rules could have been applied. We had a rule that assigned the modality of an event as HYPOTHETICAL if the expression *may* appeared in the context window. This resulted in some errors, for example,with *and there may be increased cerebral blood volume*, the modality of *volume* should be HEDGED instead of HYPOTHETICAL.

5 Conclusions and Future Work

We obtained the second best F1 score on the event polarity subtask and third best on the event span and other event attributes subtasks. We made publicly available the source code of CiTA including the rules created to produce our results. The rules to classify event and time attributes were efficient, on the other hand the list of common false positive and negative created for Run 2 did not make a significant difference.

CiTA is dependent on training data, which suggests that domain-independent approaches should be explored. One approach is to apply semantic similarity measures to automatically identify similar expressions in terms of meaning, even if using different terms (Couto and Pinto, 2013). Another approach is to explore distant supervision (Lamurias et al., 2017) to train a predictive model using a knowledge base, for example by exploring Linked Data (Barros et al., 2016), instead of annotated text.

Acknowledgments

This work was supported by the Portuguese Fundação para a Ciência e Tecnologia (https://www.fct.mctes.pt/) through the PhD Grant ref. PD/BD/106083/2015 and UID/CEC/00408/2013 (LaSIGE). We thank Mayo Clinic for providing THYME corpus and Sebastião Almeida for his contributions to our participation.

References

Marcia Barros, Francisco M Couto, et al. 2016. Knowledge representation and management: a linked data perspective. *IMIA Yearbook* pages 178–183.

Steven Bethard, Wei-te Chen, James Pustejovsky, Leon Derczynski, and Marc Verhagen. 2016. SemEval-2016 Task 12 : Clinical TempEval pages 1052–1062.

Angel X Chang and Christopher D Manning. 2012. SUTime: A library for recognizing and normalizing time expressions. *Lrec* (iii):3735–3740.

Francisco M Couto and H Sofia Pinto. 2013. The next generation of similarity measures that fully explore the semantics in biomedical ontologies. *Journal of bioinformatics and computational biology* 11(05):1371001.

Andre Lamurias, Luka A Clarke, and Francisco M Couto. 2017. Extracting MicroRNA-gene relations from biomedical literature using distant supervision. *PLoS ONE* .

Andre Lamurias, João D Ferreira, and Francisco M. Couto. 2015. Improving chemical entity recognition through h-index based semantic similarity. *Journal of Cheminformatics* 7(Suppl 1):S13.

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

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

Kaiyu Ren, Albert M Lai, Aveek Mukhopadhyay, Raghu Machiraju, Kun Huang, and Yang Xiang. 2014. Effectively processing medical term queries on the UMLS Metathesaurus by layered dynamic programming. *BMC Medical Genomics* 7(Suppl 1):S11.