ULISBOA at SemEval-2016 Task 12: Extraction of temporal expressions, clinical events and relations using IBEnt

Marcia Barros, Andre Lamurias, Gonçalo Figueiro, Marta Antunes
Joana Teixeira, Alexandre Pinheiro and Francisco M. Couto
LaSIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
mbarros@lasige.di.fc.ul.pt

Abstract

This paper describes our approach to participate on SemEval2016 Task12: Clinical TempEval. Our system was based on IBEnt, a framework to identify chemical entities and their relations in text using machine learning techniques. This system has two modules, one to identify chemical entities, and other to identify the pairs of entities that represent a chemical interaction in the same text. In this work we adapted both IBEnt modules to extract temporal expressions, event expressions and relations, by creating new CRF classifiers, lists and rules. The top result of our system was in phase2 for the identification of narrative container relations where it obtained the maximum score of precision (0.823) from all participants.

1 Introduction

In this paper we present our approach to participate on SemEval 2016 Task12: Clinical TempEval (Bethard et al., 2016) challenge. In phase1, participants had only access to the raw text, and they were asked to identify and classify time expressions (TIMEX3), event expressions (EVENT) and relations between expressions (RELATION), and in phase2 participants had access to the raw text and the manual EVENT and TIMEX3 annotations, and they were asked to identify only RELATION annotations. Our team participated in TS (identify the span of TIMEX3 expressions), ES (identify the span of EVENT expressions) and CR (identify narrative container RELATION) subtasks in phase1, and in phase2 we participated in the CR subtask. Our system was based on IBEnt (Lamurias et al., 2015), a framework that identifies chemical entities and their relations in text using machine learning techniques. Although IBEnt has been designed to extract chemical entities, we wanted to find out its potential to deal with other type of expressions. IBEnt has two modules: module one, an improvement of the tool developed by (Grego and Couto, 2013), recognizes chemical entities based on the Stanford NER software (Finkel et al., 2005) to train Conditional Random Field (CRF) classifiers using labelled data as input; module two identifies the pairs of entities that represent a chemical interaction in a given text based on machine learning techniques and domain knowledge. In particular, this latter module uses a non-linear kernel, the Shallow Language kernel (Giuliano et al., 2006) taking into account both the global and local context of each entity to determine if they are interacting or not. The Shallow Language Kernel uses the word, lemma, POS and tag of each token to train a classifier. One instance was generated for each candidate pair, whose elements were identified with a specific tag. In our system we modified both modules in order to extract the relevant expressions and relations. We used module one to extract the span of TIMEX3 and EVENT expressions, and module two to extract the RELATION between expressions. For each subtask we modify IBEnt differently. To extract TIMEX3 expression, the features of the CRF classifier were changed, as well as some specific rules and lists. To extract EVENT expressions we modified the features of the CRF classifier and we added some rules. For TIMEX3 and EVENT

Proceedings of SemEval-2016, pages 1154–1158,
San Diego, California, June 16-17, 2016.

expressions we used n-gram features, lemma, context and word shape. Identifying the RELATION between expressions required a more complex scheme, with the training of four CRF classifiers and the creations of specific rules, addressed to find relations between near entities. For RELATION we used as features the word, lemma, POS and NER tag. Thus, the aim of our system was to identify the span of temporal expressions (TIMEX3) and clinical events (EVENT), and find relations between them (RELATION).

2 Methods

To train our system we used the data set provided by Clinical TempEval organization, with 439 raw text clinical notes from Maio Clinics and the respective manual annotations, and 153 raw text clinical notes to test the system. Some adaptations were made to IBEnt framework with the intent of identifying TIMEX3 and EVENT expressions. The manual annotated data set was divided into train and development sets. We used the train set to train the CRF classifiers and the development set to evaluate and tune the performance. First, as a pre-processing step, the raw text was split into sentences using the Genia Sentence Splitter (Sætre et al., 2007).Each sentence was then processed by Stanford CoreNLP to obtain basic synthetic information to be used by the algorithms.

For TIMEX3 expressions extraction, Stanford NER already had a library for recognizing and normalizing TIMEX3 expressions, named SUTime. The problem with SUTime was that it does not recognize temporal expressions related with clinical terms, such as ”postoperative”. To solve this inconvenience, our team created a manual list, with approximately 200 temporal clinical terms, such as “post-op”, “pre-surgery” and “peritreatment”, using different combinations of words, based on temporal medical concepts from Unified Medical Language System (UMLS) (Bodenreider, 2004). To complement this list, we trained a CRF classifier with the manual annotations and we created post-processing rules to improve the results. These rules were to exclude sections that were not supposed to be annotated, such as patient medication, allergies and education; to divide dates the CRF classifier was annotating together, for example “10-06-2010 10-06-2011” thus the system could consider this as two separated entities instead of just one entity; and to exclude invalid characters, such as quotation marks, commas and parenthesis. These rules were heuristically created, based in the observation of the clinical notes raw text.

For EVENT expressions extraction we trained a CRF classifier with the training set provided. We also created rules to solve problems the classifier was not able to identify and fix. One of the rules was about the number of words an event can have (we considered that an event had just one word, which is nearly always the case). For example, “tumor demonstrated” should be considered by the system as two entities. The other rule was to exclude expressions with numbers, once the CRF classifier was annotating, for example, “T4” and “N2” as EVENT expressions.

Regarding RELATION extraction, our approach was slightly different from the ones used to extract TIMEX3 and EVENT. For this subtask, we trained four different CRF classifiers, according with the relation type: EVENT EVENT; EVENT TIMEX3; TIMEX3 TIMEX3 and TIMEX3 EVENT. Each candidate pair was considered if the two entities appear in the same or adjacent sentences. A couple of rules were also created to find relations in the raw text. First rule: if there is a TIMEX3 expression and in the next four words there is an EVENT expression, the system points a relation between both expressions (TIMEX3 and EVENT). The second rule is about the patients history: every TIMEX3 or EVENT below the EVENT HISTORY will have a relation with it. These rules were determined empirically in order to reduce the number of potential relations, since any two entities mentioned in the same document could constitute a relation.

3 Results submission

The original submitted results for phase1 had a bug. Entities, which should always have different IDs, had the same ID for EVENT and RELATION. Thus, the evaluation script provided by the organization (Chen and Styler, 2013) was unable to evaluate our system. To fix this problem, we sent the same results again but with the corrected IDs. For both

phase1 and phase2, we submitted two runs. Table1 shows the differences in our system between runs for each subtask. For TIMEX3 expressions, Phase1 Run1 (P1R1) was submitted only with the annotations obtained with the manual list of temporal concepts, without any expression obtained with the CRF classifier. On the other hand, Phase1 Run2 (P1R2) was submitted with the annotations obtained with the CRF classifier and the manual list of temporal concepts previously explained in Methods section. P1R1 and P1R2 for EVENT expressions just differ in the Stanford NER features used to train the CRF classifier, for example, MaxNGramLeng, maxLeft and useTypeSeqs. Regarding RELATION, in phase1 the main difference between runs was the input used, i.e., for this phase we used the TIMEX3 and EVENT expressions extracted from the raw text using our system. We submitted the RELATION annotated with the classifiers and with the rules (see Methods section). Phase2 submission had a greater difference between each run. Beside the fact the input was the manual annotations for TIMEX3 and EVENT provided by the organization, for Phase2 Run1 (P2R1) we submitted the RELATION obtained with either the CRF classifiers or rules, and for Phase2 Run2 (P2R2) we submitted the relations obtained with both the CRF classifiers and rules.

		TIMEX3	EVENT	RELATION
Phase1 with bugfix	Run1(P1R1)	List	Features	TIMEX3 and EVENT from P1R1
	Run2(P1R2)	List and CRF classifier	Features	TIMEX3 and EVENT from P1R2
Phase2	Run1(P2R1)	-	-	CRF classifier or Rules
	Run2(P2R2)	-	-	CRF classifier and Rules

Table 1: Main differences between run1 and run2 for phases 1 and 2. List: list of manually curated temporal expressions; Features: optimized features for precision/F1-measure; CRF classifier: CRF classifier trained with different training sets.

4 Results and Discussion

In this section we exhibit the results of ULISBOA for the participating subtasks and we discuss these results. In Table 2 we present ULISBOA results for both phases 1 and 2, and the maximum scores obtained in the competition for each subtask.

4.1 TS identifying the spans of time expressions(TIMEX3)

For TIMEX3 expressions extraction there is a major difference in the results of P1R1 and P1R2. For this subtask we intended to modify Stanford NER features, focusing P1R1 for a better recall and P1R2 for a better precision. Instead, and due to a misunderstanding, P1R1 was submitted only with the TIMEX3 annotations obtained through the manual list of temporal concepts, which justifies the low recall. However, after a deeper analysis of this problem, we concluded that precision should be higher, once it was assumed that all the terms in the list are always a TIMEX3 expression. With a posterior revision of the list (after the deadline of submission), we tested it again. Terms included in the list, like "at this time" and "at that time", were removed and we added all terms that were already in the list with the first letter in uppercase, with this obtaining a higher precision (0.933). This test suggest that the temporal list we have created is a good complement to our system. For P1R2, the score achieved is near the maximum for precision (-0.064), recall (-0.066) and F1-measure (-0.063). There are some words our system did not annotate as TIMEX3 expression, for example, "time" and "date". We also observed that there was some ambiguity on the annotations regarding the spans of expressions with more than one words. For example, there were cases where "at this time" was annotated, and others where only "this time" was annotated. These inconsistencies made it more difficult to train a CRF classifier and to generate a list of TIMEX3 expressions. Another issue was about our manual list of temporal concepts, which did not consider uppercases words, so the system did not annotate terms such as "Intraop" as TIMEX3.

4.2 ES identifying the spans of event expressions(EVENT)

P1R1 and P1R2 scores were similar in EVENT expressions extraction, with a difference between runs of 0.002 for precision, 0.006 for recall and 0.004 for F1-measure. For both runs we used different Stanford NER features, not achieving significant differences in results. However, P1R1 was the closest to the maximum score, especially regarding to precision (-0.034). Once this differences are not statisti-

		TIMEX3			EVENT			RELATIONS		
		P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Phase1 with bugfix	Run1(P1R1)	0.623	0.065	0.118	0.881	0.745	0.807	0.122	0.009	0.017
	Run2(P1R2)	0.776	0.692	0.732	0.879	0.739	0.803	0.108	0.009	0.017
	Max	**0.840**	**0.758**	**0.795**	**0.915**	**0.891**	**0,903**	**0.531**	**0.471**	**0.479**
Phase2	Run1(P2R1)	-	-	-	-	-	-	0.273	0.255	0.264
	Run2(P2R2)	-	-	-	-	-	-	**0.823**	0.056	0.105
	Max	**-**	**-**	**-**	**-**	**-**	**-**	**0.823**	**0.564**	**0.573**

Table 2: ULISBOA results for phase1 (TIMEX3, EVENT and RELATION) and phase2 (RELATION). P=precision; R=Recall; F1=F1-measure; Max=maximum score obtained in the challenge.

cally significant, we are going to focus our discussion in P1R1. For this subtask, we trained a CRF classifier and combined the results with the rules explained in Methods sections. We did not use any ontology or dictionary which explains why some basic terms were not annotated (False Negatives), such as "scan" and "normal", or which were incorrectly annotated (False Positives), such as "CT-scan" and "plan". Our system achieved a score of 0.80 to F1-measure in the aforementioned subtask, i.e., a considerable part of the expressions was correctly classified as EVENT. However, classifying an expression as EVENT depends on its surrounding context. In the previous example, "scan", as the action of doing an exam, should be annotated as an EVENT, and "CT-scan", as X-Ray Computed Tomography machine, should not be annotated. Thus, there is an increased demand to develop new semantic techniques, such as semantic similarity, to complement our system (Couto and Pinto, 2013). Another problem was about the rule that considers the EVENT only as one word, because instead of separate the entity in two different entities, this rule excluded these entities from the results, so we were losing one entity and lowering the recall.

4.3 CR identifying narrative container relations(RELATION)

For Clinical TempEval phase1, our system RELATION results were much lower than the maximum scores, for both P1R1 (precision -0.409; recall -0.471; F1-measure -0.462) and P1R2 (precision -0.423; recall -0.471; F1-measure -0.462). The variances in the system between each run in phase1 were only the input, TIMEX3 and EVENT expressions extracted from the raw text with our system.

In phase2, our results were improved when compared with phase1 (see Table2). For example, in phase1, the same clinical note had 3 RELATIONS identified, compared with 109 RELATIONS identified in phase2. A better recall was obtained in phase2, for both runs. For P2R1 we considered relations identified either by the classifiers or rules, which is why we got a better recall for this run. For P2R2, we considered relations identified both by the classifiers and rules, i.e., we only accepted relations as positive if the same relation was identified by both the classifier and the rules. For example, in the sentence "the oncologist today would attempt to quantify", in P2R1 the relation between "today" and "quantify" was detected by the rules, but not by the classifier, showing in the final results as a false positive. In P2R2 this did not happen, being this relation excluded from final results because only the rules had detected it. In P2R2 our system achieved a maximum score of 0.823 for precision. However, the recall was lower than P2R1, since we limited the number of relations to be considered. Thus, we can affirm that our system in P2R2 found less relations, but with much higher precision.

5 Conclusions and Future Work

In SemEval Clinical TempEval we participated in three of the six subtasks: TS (TIMEX3 span extraction), ES (EVENT span extraction) and CR (RELATION extraction). Our system, based on machine learning and rules, achieved the maximum precision score for RELATIONS in Phase2. However, more work is necessary to understand how we can improve the recall without a significant decrease in precision. Despite the fact that we had limited time to work for this challenge, our scores for TIMEX3 and EVENT expressions extraction were acceptable and near the maximum scores achieved for other

systems. For future work, we intend to correct errors already found and test the system again, and for EVENT expressions we want to test several ontologies, such as SNOMED-CT (Cornet and de Keizer, 2008), to improve our results. Due to the ambiguity of clinical notes, we must understand the real impact manual rules have in final results, so we may have the opportunity to create a greater number and more accurate rules in order to be applied in a larger number of cases. One technique that can be used is distant learning (Mintz et al., 2009). This technique uses a knowledge base, such as an ontology, to automatically generate training data, requiring less manual effort. Another approach is to extend our domain knowledge about the problem in hand by for example collecting and exploring semantic web medical resources (Machado et al., 2015). Furthermore, we want to create an open source framework for our system, after we make it stable. Next SemEval Clinical TempEval edition, we hope to participate in the remaining subtasks and to improve our results, especially for RELATION, where much work remains to be done.

Acknowledgments

This work was supported by FCT through funding LaSIGE Research Unit, ref. UID/CEC/00408/2013.

References

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. Semeval-2016 task 12: Clinical tempeval. June.

Olivier Bodenreider. 2004. The unified medical language system (umls): inte-grating biomedical terminology. *Nucleic acids research*, 32(1):267–270.

Wei-Te Chen and Will Styler. 2013. Anafora: A web-based general purpose annotation tool. pages 14–19, June.

Ronald Cornet and Nicolette de Keizer. 2008. Forty years of snomed: a literature review. *BMC medical informatics and decision making*, 8(1):S1:S2.

Francisco M Couto and 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.

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*, pages 363–370.

Claudio Giuliano, Alberto Lavelli, and Lorenza Romano. 2006. Exploiting shallow linguistic information for relation extraction from biomedical literature. *EACL*, 18:401–408.

Tiago Grego and Francisco M. Couto. 2013. Enhancement of chemical entity identification in text using semantic similarity validation. *PloS one*, 8(5):e62984.

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

Catia Machado, Dietrich Rebholz-Schuhmann, Ana Freitas, and Francisco M Couto. 2015. The semantic web in translational medicine: current applications and future directions. *Briefings in bioinformatics*, 16(1):89–103.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. *Association for Computational Linguistics*, 2.

Rune Sætre, Kazuhiro Yoshida, Akane Yakushiji, Yusuke Miyao, Yuichiro Matsubayashi, and Tomoko Ohta. 2007. Akane system: protein-protein interaction pairs in biocreative2 challenge, ppi-ips subtask. In *Proceedings of the Second BioCreative Challenge Workshop*, pages 209–212.