KULeuven-LIIR at SemEval 2016 Task 12: Detecting Narrative Containment in Clinical Records

Artuur Leeuwenberg and **Marie-Francine Moens**
Department of Computer Science
KU Leuven, Belgium
{tuur.leeuwenberg, sien.moens}@cs.kuleuven.be

Abstract

In this paper, we describe the KULeuven-LIIR system at the Clinical TempEval 2016 Shared Task for the narrative container relation sub-task (CR). Our approach is based on the cTAKES Temporal system (Lin et al., 2015). We explored extending this system with different features. Moreover, we provide an error analysis of the submitted system, and report on some additional experiments done after submission.

1 Introduction

We describe the KULeuven-LIIR submissions for the Clinical TempEval 2016 shared task (Bethard et al., 2016). Our motivation for this first participation is to gain insight into the task, and the data, as a basis for future work. We participated in the narrative container relation (CR) task. In the CR task narrative containment relations between events, and events and temporal expressions are to be extracted. Two examples of such relations are given in Sentence 1.

(1) A *colonoscopy* on *27 September 2008* revealed a circumferential *lesion*.

The relations that are to be extracted are

- CONTAINS(*27 September 2008*, *colonoscopy*)
- CONTAINS(*colonoscopy*, *lesion*)

In the shared task, the clinical records on colon cancer from the THYME corpus are used (Styler IV et al., 2014). We participated in phase 2 evaluation, which means that manually annotated EVENT and TIMEX3 spans with their attributes are provided as part of the input. We submitted two runs that are both based on the cTAKES-Temporal system (Lin et al., 2015). For the first run, we used cTAKES-Temporal with a set of already provided features. We further explain the cTAKES-Temporal system and the features that we used in Section 2. For the second run, we added new features, that are based on an error analysis of a small sample from the results of running the first system on the development set. These additional features are described in Section 3. Moreover, we describe a more elaborate error analysis of the second system in Section 4, and some experiments that we performed with the aim to handle these errors in Section 5.

2 The cTAKES Temporal System (run 1)

The first version of the temporal module in cTAKES[1], an extensive open source information extraction system for clinical free-texts (Savova et al., 2010), is described by (Lin et al., 2015). We used the currently available version in our experiments. The current version of cTAKES Temporal consists of two SVM classifiers to detect containment. One for containment between events (EE), and one for containment between temporal expressions and events (TE), each using different features, as the nature of the two types of relations is quite different. The features we used for each classifier are described in Table 2, and will be explained further in the next paragraph. Note that with entity we refer to either an EVENT or TIMEX3 expression, which can con-

[1] ctakes.apache.org/

Proceedings of SemEval-2016, pages 1280–1285,
San Diego, California, June 16-17, 2016.

sist of several tokens (words, numbers, punctuation etc.). The search space, or the entity-pairs that the system considers are all EE and TE pairs in each sentence or line, in each document for their respective classifier (the EE SVM, or the ET SVM).

As in the THYME corpus only the heads of events are annotated, cTAKES-Temporal expands these heads to their full phrase (e.g., the annotated head *cancer* expands to *ascending colon cancer*). Afterwards it creates extra training data by not only considering the pairs of entity phrases, but also the sub-phrases (e.g., *cancer*, *colon cancer*, and *ascending colon cancer*).

The precision, recall, and F-measure that our first submission scored on the test set is reported in Table 5. What can be noticed from the top-part of Table 5 is that the cTAKES Temporal system (run 1) gives reasonable results, scoring ∼9% higher than the median F-measure of all systems evaluated in Clinical TempEval 2016, and ∼4% short compared to the best system.

2.1 Features

The features used in run 1 correspond to those described by (Lin et al., 2015), and feature extraction is included within the (open source) cTAKES Temporal module. In this module, tokenization is done following the Penn Treebank rules (Mott et al., 2009). Sentence boundaries, parts-of-speech, chunks, and constituency parses (for finding heads) are obtained using the corresponding cTAKES modules. Dependency paths are extracted using the ClearNLP dependency parser model in cTAKES. UMLS types are extracted by means of direct token or chunk matches, and by creation of lexical variants using the UMLS Lexical Variant Generation package [2].

3 Additional Features (run 2)

Our motivations for the additional features come from analyzing a small sample of errors (small due to time constraints at the moment of submission) of the first system on the development set, from which we concluded that especially long distance and sometimes also short distance relations characterized by prepositions (especially *of* and *with*) were not found by the system. An example of such a relation missed by the system is CONTAINS(*colonoscopy*, *polypectomy*) in the fragment given in Figure 3.

The first added feature is a modified version of the dependency path (MDP). We include part of speech tags in the path for nouns, personal pronouns, numbers, adjectives and determiners. For other parts of speech we include the words themselves. In the original dependency path feature of cTAKES Temporal, only parts of speech are included. The second

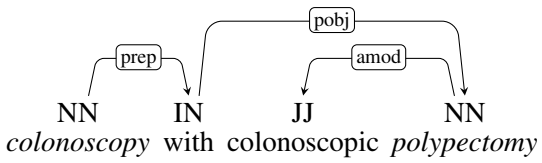

Figure 1: Sentence fragment, with its POS, and dependencies.

feature consists of the sequence in-between the two entities (IBS), i.e., the words in-between connected by an underscore. We replace the same categories as for the MDP by their respective POS in the sequence. The modified dependency path, and the in-between sequence (IBS) feature for the small fragment in Figure 3, where we consider *colonoscopy*, and *polypectomy* are:

MDP: `<NN>prep<with>pobj<NN>`

IBS: `with_JJ`

The precision, recall, and F-measure on the test set for the second run are shown in Table 5. Adding these features resulted only in a tiny increase in both precision and recall (0.001). From this we conclude that our error sample might not have been representative enough for the rest of the corpus, or that the features do not cover the errors made. Also, there is probably overlap between the MDP feature and the existing dependency path feature, and between the IBS feature and the existing Tokens feature explaining why adding the MDP and IBS features do not help much.

4 Error Analysis

4.1 Error Categories

We also ran the system of the second run on the development set of the THYME corpus (but now only trained on the train set) and analyzed its errors with

[2] https://www.nlm.nih.gov/research/umls

Feature	Description	Extraction
1. Tokens	String features for the tokens within each entity, the entity's first and last token, and the tokens in-between the two entities.	cTAKES-Temporal module, with Penn Treebank tokenization rules (Mott et al., 2009)
2. POS	String features for the parts of speech of the tokens of both entities.	cTAKES' POS tagger
3. Event Attributes	String features for each event modality, polarity, and event type of both entities.	Part of the input (phase 2 evaluation)
4. UMLS	Boolean flags for the semantic UMLS type of each entity, and as a paired feature ($type_1$-$type_2$).	cTAKES-Temporal module, with UMLS Lexical Variant Generation
5. Dependency Path	The dependency path between the entities, comprised of edge-labels, and POS tags.	cTAKES dependency parser (clinical ClearNLP model)
6. Overlapped Head	Consists of a string feature for each overlapping head word, and numerical features for the count and ratio of overlapping words.	cTAKES Temporal module, where heads are found by means of the cTAKES Constituency Parses
7. Nearest Flag	Boolean flag on if two entities are the closest candidates of each other.	cTAKES Temporal module
8. Conjunction Flag	Boolean flag for conjunctions in-between the entities only *CC*, *COMMA*, or *IN* POS tags are considered.	cTAKES Temporal module (using cTAKES's POS tagger)
9. Special Word Relation	Boolean feature for special phrase types in-between the entities (e.g., *starts out*: STARTING).	Manually constructed list by (Lin et al., 2015)
10. Temporal Attributes	String features for EVENT modality, and TIMEX3 class.	Part of the input (phase 2 evaluation)

Table 1: Feature descriptions for the EVENT-EVENT classifier, using features 1, 2, 3, 4, 5, 6 and the TIMEX3-EVENT classifier, using features 1, 3, 5, 7, 8, 9, 10.

category	count	percentage
Correct	22	44%
Questionable	10	20%
Negation	5	10%
Other	13	26%

Table 2: Categories of **false positives** for the run 2 system on the THYME development set.

regard to predicting the correct CONTAINER relations. Around 69% of the errors were false negatives, and around 31% false positives. We then randomly sampled 50 errors of both types and categorized them ($\sim$1.5 % of false negatives, and $\sim$3.3% of false positives). For the false positives, we found four frequent categories: (1) predictions that we judge to be *correct* positive predictions, (2) predictions that we judge the correctness to be *questionable*, (3) incorrect predictions caused by an unprocessed *negation*.

In Table 2 the statistics of the categories are shown. What can be noticed is that many of the false positives are questionable or correct. This indicates that the actual precision of the system might be higher. On the other hand, the consistent errors that the system makes result from ignoring negations or negating verbs (e.g., *to quit*). The errors in the category 'other' had various causes, such as wrong tokenization, mistakes in event annotation, or sometimes the cause was unclear.

A bigger, and maybe more interesting, source for improvement are the false negatives, i.e., the relations missed by the system. Here we found the following major (non-mutually exclusive) categories: (1) *cross-sentence* relations, (2) *unknown tokens (UNKS)* are tokens that appear in the dev-set as argument of the relation, but not in the train-set. Categories (1) and (2) appear to be a common source of false negatives. Another frequent error category is (3) the relation *crosses a newline*[3], (4) the relation is part of a data *table*, as shown in Figure 2, (5) one of the arguments of the relation is part of a within-sentence enumeration (usually comma-separated), (6) some false negatives we judge to be *correctly* predicted as negative. The proportions of these categories are shown in Table 3. In our notion of UNK errors, we exclude unseen dates (e.g.,

[3]Note that sentence boundaries are given by the sentence boundary detection, and new lines by the newline token ('\n').

category	count	percentage
Cross-sentence	13	26%
UNKs (excl. dates)	13	26%
Cross-newline	11	22%
Table	9	18%
Enumeration	6	12%
Correct	4	8%
Other	10	20%

Table 3: Categories of **false negatives** for the run 2 system on the THYME development set.

February 2, 2008) as they include digits, resulting in many (rather predictable) UNK tokens.

4.2 Examples of Errors

The UNK tokens from the false negatives seem a mix of more domain specific and general terms. More domain-specific terms are *CK5-6*, *seepage*, *mesh*, and *Mumps*. The more general are *outline*, *QUALITY*, *function*, and *question*. The cross-newline and table error categories have some overlap. An example of the cross-newline category is shown in Figure 2, where the missed relation in our random sample is CONTAINS(*HISTORY*, *affected*). But also the containment between *HISTORY* and *Polio*, *Obesity*, *equivalency*, *hyperlipidemia*, *Hypertension*, *PVD*, *stenosis*, and *disease* are missed by the system. A table-error would look similar, but with test measurement reports, such as for example patient height, weight, or body mass index.

```
PAST MEDICAL HISTORY
1) Polio at age 15. RUE affected.
2) Obesity.
3) CAD equivalency with hyperlipidemia ...
4) Hypertension, well-controlled.
5) PVD of left leg. Stable and asymptomatic.
6) Moderate aortic stenosis.
7) Non-alcoholic fatty liver disease.
```

Figure 2: An example of the cross-newline error category, where the CONTAINS(*HISTORY*, *affected*) relation is missed by the system (and many more).

The enumeration error category contains within sentence enumerations. An example sentence is given in sentence (2).

(2) We have discussed the *characteristics* of the cancer, including the type of malignancy, size, grade, lymph node *status*, and stage of the cancer.

Here, CONTAINS(*characteristics*, *status*) is the missed error in our sample. One reason for this type of errors might be that the currently used dependency parser chains the parts of such enumerations, resulting in a long dependency path between the two entities, which is less likely to occur frequently in the training data.

4.3 Token Frequency

We also looked at token diversity in the training set for event-event containment, and time-event containment. What can be noticed from Table 4 is that there are slightly more EE container relations, than TE container relations. It is also striking that the source (first argument of the containment relation, i.e., the container) of the TE relations have a relatively low average token frequency, and a high percentage of UNKs, compared to the source of the EE relations. The explanation for this is that for the TE pairs the source category (TIMEX3) of tokens includes dates, containing numbers, causing big token diversity (each date is a new token). To get a better idea of token diversity without considering each number as a different token we conflated the digits in TE source tokens (e.g., *February 2, 2008* becomes *February 5, 5555*). In-between the brackets the percentage of UNKs is shown, after digit conflation of these tokens. Furthermore, the source of both the TE and the EE relations, i.e., the container, seems to be less diverse compared to the target, i.e., the event that is contained. We conclude that this is because of the relatively low vocabulary size of this category, and low percentage of UNKs, i.e., there are less different containers (container EVENT or TIMEX3 expressions) than containees (contained EVENTS).

5 Out-of-Competition Experiments

We conducted two sets of experiments, one set to address the UNK-related errors, and one to address the cross-newline errors.

To tackle the UNK token category, without adding extra resources, we experimented with adding subword features, in particular character 3-grams of the entities. We also experimented with adding word embeddings to the event-event SVM. We trained

	Event-Event		Time-Event	
	source	target	source	target
|train-tokens|	5931	5931	5093	5093
|train-vocabulary|	637	1417	1268 (839)	1198
avg. token-freq.	9.3	4.2	4.0 (6.0)	4.3
UNK percentage	0.10	0.18	0.43 (0.17)	0.15

Table 4: Token-frequency statistics. The 'source' of the containment relation refers to the container, and taget to the containee. For TE sources we also calculated statistics after digit conflation, shown in-between brackets.

the embeddings on 10 million words of crawled web data from the oncology/colon cancer domain[4] plus the THYME training data. To construct the word vectors we used the continuous bag-of-words (CBOW), and Skip-gram model (SG) by (Mikolov et al., 2013). The vectors are of dimension 250, and were trained using the default settings in the word2vec toolkit[5]. What can be noticed from Table 5 is that adding these features did not significantly improve the F-measure.

System	P	R	F
clinical tempeval 2016 – median	0.589	0.345	0.449
clinical tempeval 2016 – best	**0.823**	**0.564**	**0.573**
run 1	0.714	0.428	0.536
run 2	0.715	0.429	0.536
run 2 + character 3-grams	0.706	0.428	0.533
run 2 + CBOW^{EE}	0.701	0.438	0.539
run 2 + SG^{EE}	0.708	0.433	0.537
run 2 + newline extension	**0.714**	0.441	0.545
run 2 + newline ext. + CBOW^{EE}	0.706	0.452	0.551
run 2 + newline ext. + CBOW^{EE} + SG^{EE}	0.705	**0.452**	**0.551**

Table 5: THYME test set performance of different system settings. The lower part of the table comes from experiments done out-of-competition.

With the goal to handle errors caused by separation by a newline, we extended the search space by adding candidate pairs (either EE, or ET) that were separated by a new line, but for which the end of the line consisted of a comma, or colon symbol. This increased recall (by 2%), and thus F-measure slightly as well (by 1%). This small increase in performance does not match with the relative size of the cross-newline error category, so there is still much room for improvement.

[4] emedicine.medscape.com, Pubmed Central & Pubmed

[5] code.google.com/archive/p/word2vec/

6 Conclusions & Future Work

In this paper we described the first participation of KULeuven-LIIR in the Clinical TempEval Shared Task 2016 (Bethard et al., 2016). Our motivation to participate was to gain insight into the CONTAINER relation extraction task, and the data for future research. We started from the cTAKES-Temporal system (Lin et al., 2015) and experimented with adding various features. From our experiments we conclude that adding the modified dependency path feature, or in-between sequence features does not improve the performance of the system. Furthermore, from our cTAKES Temporal module, we noticed that cross-sentence relations, cross-newline relations, and UNKs are an important source of error. Also, false negatives seem a bigger problem than false positives. To tackle the problem of UNKs, we experimented with adding word embeddings, trained on in-domain web crawled text, as a feature, and adding a character n-gram feature. These features did not seem to contribute significantly. For future research it could be interesting to explore alternative methods to deal with UNKs. To tackle the cross-newline mistakes the system made, we extended the scope of the candidate pair creation across new lines, which improved recall on the test set slightly, but still leaves much room for improvement.

Acknowledgments

The authors would like to thank the reviewers for their constructive comments which helped us improve the paper. Also, we would like to thank the Mayo Clinic for permission to use the THYME corpus. This work was funded by the KU Leuven C22/15/16 project "MAchine Reading of patient recordS (MARS)", and by the IWT-SBO 150056 project "ACquiring CrUcial Medical information Using LAnguage TEchnology" (ACCUMULATE).

References

Steven Bethard, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2014. Clinical TempEval. *arXiv preprint arXiv:1403.4928*.

Steven Bethard, Leon Derczynski, Guergana Savova, Guergana Savova, James Pustejovsky, and Marc Ver-

hagen. 2015. Semeval-2015 task 6: Clinical Tempeval. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. SemEval-2016 task 12: Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

Quynh Thi Ngoc Do, Steven Bethard, and Marie-Francine Moens. 2015. Domain adaptation in semantic role labeling using a neural language model and linguistic resources. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(11):1812–1823, November.

Chen Lin, Dmitriy Dligach, Timothy A Miller, Steven Bethard, and Guergana K Savova. 2015. Multilayered temporal modeling for the clinical domain. *Journal of the American Medical Informatics Association: JAMIA*.

Alexa T McCray, Suresh Srinivasan, and Allen C Browne. 1994. Lexical methods for managing variation in biomedical terminologies. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 235. American Medical Informatics Association.

Tomas Mikolov, Stefan Kombrink, Anoop Deoras, Lukar Burget, and Jan Cernocky. 2011. RNNLM-Recurrent neural network language modeling toolkit. In *Proc. of the 2011 ASRU Workshop*, pages 196–201.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Justin Mott, Ann Bies, Colin Warner, and Ann Taylor. 2009. Supplementary guidelines for ETTB 2.0. *University of Pennsylvania*.

Guergana K Savova, James J Masanz, Philip 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. *Journal of the American Medical Informatics Association*, 17(5):507–513.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.