KULeuven-LIIR at SemEval-2017 Task 12: Cross-Domain Temporal Information Extraction from 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 system of the KULeuven-LIIR submission for Clinical TempEval 2017. We participated in all six subtasks, using a combination of Support Vector Machines (SVM) for event and temporal expression detection, and a structured perceptron for extracting temporal relations. Moreover, we present and analyze the results from our submissions, and verify the effectiveness of several system components. Our system performed above average for all subtasks in both phases.

1 Introduction

In this paper, we describe the system used for the KULeuven-LIIR submissions at SemEval task 12, named Clinical TempEval 2017 (Bethard et al., 2017), which is concerned with temporal information extraction from clinical records. In Clinical TempEval extraction of temporal information is split into six subtasks. Our system participated in all tasks:

1. Detection of event spans (ES)
2. Identification of event attributes (EA)
3. Detection of temporal expressions (TS)
4. Attribute identification of temporal expressions (TA)
5. Extraction of document-creation-time relations for events (DR)
6. Extraction of narrative container relations (CR)

This year, a new aspect of Clinical TempEval is that systems will be evaluated across domains, which involves two phases: Firstly, *unsupervised domain adaptation* (Phase I), where the training data is in the colon cancer domain, and the test data in the brain cancer domain. And secondly, *supervised domain adaptation* (Phase II), where the vast majority of the training data are colon cancer reports, and a small number of brain cancer reports is made available for training as well. The test data is again in the brain cancer domain.

Our system consist of a combination of linear Support Vector Machines (SVM) for entity span and attribute recognition (tasks ES, EA, TS and TA), and a document-level structured perceptron (Leeuwenberg and Moens, 2017) for relation extraction tasks (tasks DR and CR). We used three system components for the domain adaptation: (1) assigning more weight to target-domain training data, (2) introduction of a UNK (unknown) token to model out-of-vocabulary words, and (3) exploitation of relational properties of temporality during prediction.

In Section 2, we provide a detailed description of our full system, and in Section 3 we discuss the results from our submissions.

2 Our System

Our system consist of three main components (1) preprocessing, (2) entity detection, and (3) relation extraction. In Figure 1, we show a schematic overview of our system.

2.1 Preprocessing

The corpus used in Clinical TempEval 2017 is the THYME corpus (Styler IV et al., 2014). For the unsupervised domain adaptation phase (Phase I), we use all colon cancer sections for training. For the supervised domain adaptation (Phase II) participants also received a small training section in

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 1030–1034,
Vancouver, Canada, August 3 - 4, 2017.

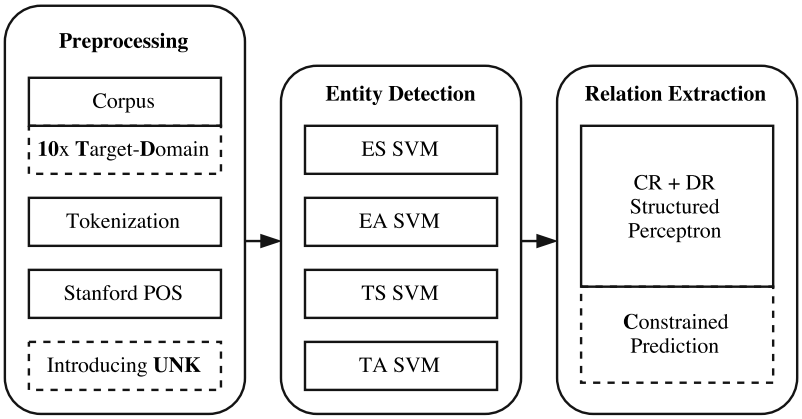

Figure 1: Schematic overview of our system. Components we expect to help domain adaptation are dashed.

the brain cancer domain. Some statistics about the dataset can be found in Table 1.

Table 1: Dataset statistics for the THYME sections used in our experiments.

Section	Documents
Training Colon Cancer	591
Training Brain Cancer	30
Test Brain Cancer	148

Our first simple method for adapting to a new domain, when given target-domain training data (Phase II), is to assign more weight to the target-domain data at training time (Jiang and Zhai, 2007). In our submissions we assigned a 10 times higher weight to the target-domain training data compared to the colon cancer training data.

In all experiments, we preprocess the text by using a very straightforward tokenization procedure considering punctuation[1] or newline tokens as individual tokens, and splitting on spaces. We also employ lowercasing, and conflate all digits to a single representation. An example would be:

October 20, 1991 ⇒ *october 55 , 5555*

For our part-of-speech features, we rely on the Stanford POS Tagger (Toutanova et al., 2003), with the English bidirectional tagger model. We also take the transitive closure of the CONTAINS relation on the training data, as this has shown to improve results in existing work (Mani et al., 2006).

[1], . / \ " ' = + − ; : ( ) ! ? < > % & $ * | [ ] { }

Our second domain adaptation modification involves the introduction of an unknown word token (UNK) to the input vocabulary of the extraction models. This is a widely used technique in statistical language modeling to account for out-of-vocabulary (OOV) words. In a language modeling setting, we can expect that the proportion of OOV words in the test set can be modeled by using the proportion of one-time-occurring words from the training set, by Good Turing estimation (Gale and Sampson, 1995). In our system, we train the weights for the UNK token by replacing all tokens that occur only once in the training data by the UNK token. At prediction time we simply replace all words that are OOV by the UNK token. We expect this technique to be effective for domain adaptation as new words can be a serious problem when crossing domains.

2.2 Entity Detection

For all span and attribute tasks we employ linear SVM classifiers[2]. We only resort to token and POS features, and use the same features for span detection as for attribute detection. More elaborate feature descriptions are shown in Table 2. We consider all single tokens as EVENT candidates, and all token {1,2,3,4,5,6}-grams as TIMEX3 candidates (upper bound 6 is based on tuning on the colon cancer training data).

2.3 Relation Extraction

For relation extraction we rely on the linear document-level structured perceptron by

[2]Trained using LIBLINEAR(Fan et al., 2008) with regularization constant C=1.0 (tuned on the colon cancer section of the training data from {0.1, 1.0, 10})

Table 2: Features of the local feature functions of each subtask: ϕ_{cr} for CR, ϕ_{dr} for DR, ϕ_{e*} for ES and EA, and ϕ_{t*} for TS and TA.

Features	ϕ_{dr}	ϕ_{cr}	ϕ_{e*}	ϕ_{t*}
Strings for tokens and POS of each entity	✓	✓	✓	✓
Strings for tokens and POS in a window of size {3, 5}, left and right of each entity	✓	✓	✓	✓
Booleans for entity attributes (event polarity, event modality, event degree, and type)	✓	✓		
Strings for tokens and POS of the closest verb	✓			
Strings for tokens and POS of the closest left and right entity	✓			
Strings for token {1, 2, 3}-grams and POS {1, 2, 3}-grams in-between the entities		✓		
Booleans on if the first argument occurs before the second (w.r.t. word order)		✓		

Table 3: Global DR (document-level) features.

Feature	**Description**
Φ_{sdr}	Bigram and trigram counts of subsequent DCTR-labels in the document

Leeuwenberg and Moens (2017)[3]. Their model employs a structured learning paradigm, assigning a score S to each label assignment. Prediction corresponds to finding the label assignment with the highest score. The score for a document-level label assignment is constructed by joining all local features (shown in Table 2) within a document for both tasks (DR and CR), together with a global DR feature shown in Table 3, resulting in a joint feature vector $\Phi(X, Y)$.

The joint features $\Phi(X, Y)$ are assigned a weight vector λ, resulting in the linear scoring function in Equation 1.

$$S(X, Y) = \lambda\Phi(X, Y) \tag{1}$$

The weight vector λ is trained using the structured perceptron algorithm (Collins, 2002), with averaging (Freund and Schapire, 1999).

At prediction time integer linear programming (ILP) is used to find the best label assignment Y^*, as shown in 2, using the Gurobi ILP Solver (Gurobi Optimization, 2015).

$$Y^* = \arg\max_{Y} S(X, Y) \tag{2}$$

We also experimented with the constraints on the output labeling formulated by Leeuwenberg and Moens (2017). The constraints enforce the model to output labeling to be temporally consistent, by enforcing relational properties onto the predictions. We only chose the properties relevant for the CR and DR subtasks, which are transitivity of containment, but also consistency between containment and the document-creation time relations of the events. The relational properties that we enforce as constraints during prediction are captured in the following rules (condition above, and conclusion below the horizontal line):

$$\frac{\text{contains}(x, y) \wedge \text{contains}(y, z)}{\text{contains}(x, z)} \tag{3}$$

$$\frac{\text{contains}(x, y) \wedge \text{before}(x, doctime)}{\text{before}(y, doctime)} \tag{4}$$

$$\frac{\text{contains}(x, y) \wedge \text{after}(x, doctime)}{\text{after}(y, doctime)} \tag{5}$$

Our hypothesis is that these constraints can help with assigning labels to unfamiliar input (e.g. from the target-domain), by ensuring that local assignments are consistent with surrounding labels.

3 Experiments and Results

We conducted a number of experiments with our system to test the effectiveness of the different system components[4]. We submitted in phase I, and in phase II. In Phase I, we used our system as shown in Figure 1, only with the UNK introduction, so without increased weight for target-domain training data (as there is none in Phase I), and without constraints. In Phase II, our system includes the full system, with all proposed components for domain adaptation.

When we look at the results of Phase I, in Figure 2, we can see that our system performs above average on all tasks, and for both attribute identification tasks, it performs best (for EA there is another system with best performance).

If we look at the results of Phase II, in Figure 2, our system again performs above average in all

[3]Using the code at https://github.com/tuur/SPTempRels

[4]Code at https://github.com/tuur/ClinicalTempEval2017

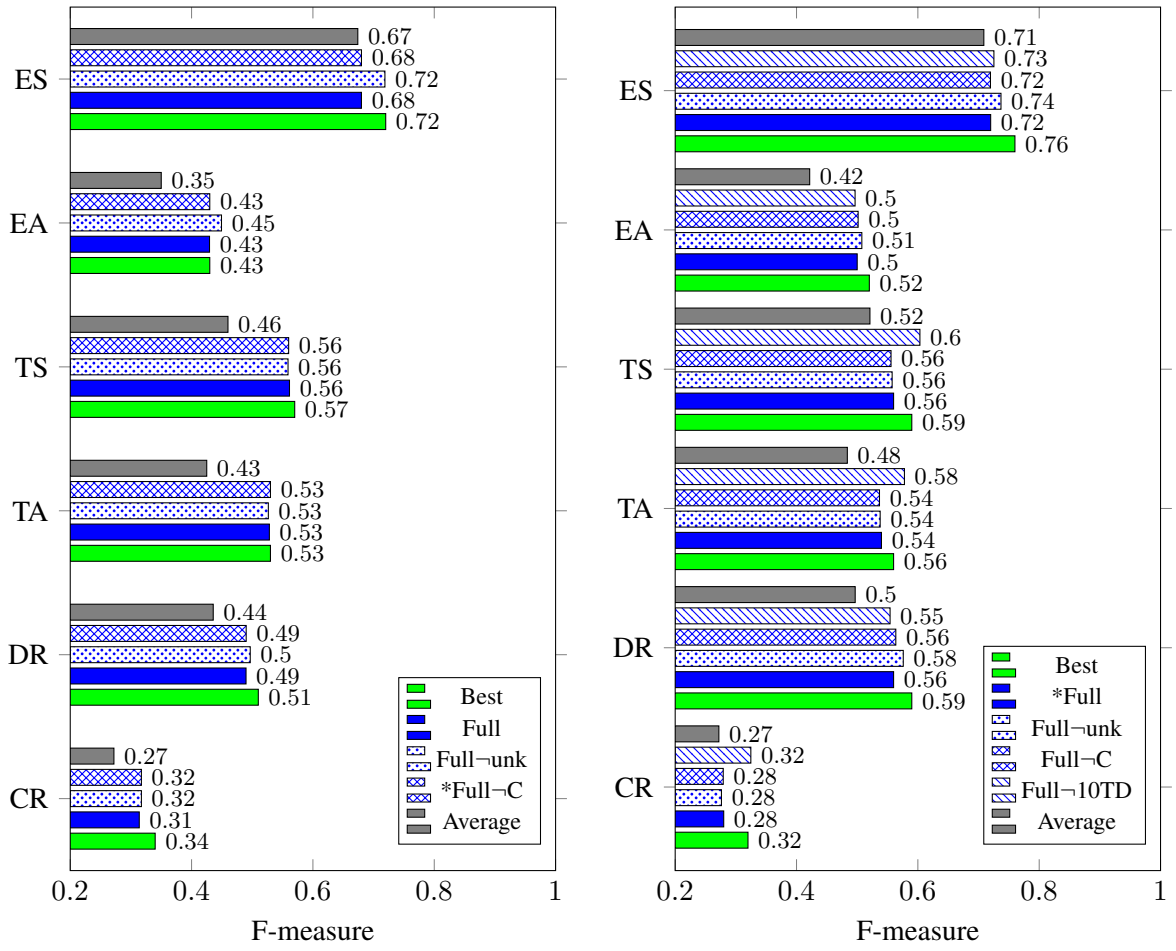

Figure 2: Results from Phase I (left) and Phase II (right): We compare our submission (indicated by *) to the best performing system, and to the average score of all participating systems in each task. Out of competition, an ablation of each modification is also evaluated (¬ indicates absence of a component).

cases. However, it seems our system does not lie as close to the best system as in Phase I, suggesting that we could have better exploited the target-domain training data.

When looking at the ablation of the system components in Figure 2, we can see that using the UNK modification (comparing Full with Full¬unk), decreases performance for the ES, EA and the DR subtask. Furthermore, employing temporal constraints (C) appears to have a slightly negative influence in Phase I for DR, and little influence in Phase II.

The effect of adding more weight to target-domain training data (10TD) is mixed, leaning towards a negative influence. For DR performance increased by 1 point (because of increase in precision). However, for CR, TS, TA and EA it seems to have a negative effect, for various reasons. For example, for CR mostly due to a big decrease in recall, but for TS due to a big decrease in precision (hardly any difference in recall). This shows that the effectiveness of weighting the target-domain training data is highly task-dependent.

An interesting observation is that there is hardly improvement in CR performance in Phase II compared to Phase I (the best system score is even lower). This suggests that domain-adaptation for CR is more challenging than the other subtasks.

4 Conclusions

We described the KULeuven-LIIR system at Clinical TempEval 2017, for all six subtasks. Our system exploits SVM for entity detection and a document-level structured perceptron for relation extraction. Our system performed above average for all subtasks in both phases. For future research it would be interesting to analyze the errors that were made by the system, and explore methods to better exploit small amounts of target-domain training data, or unlabeled target-domain data.

Acknowledgments

The authors would like to thank the reviewers for their constructive comments, and 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, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. Semeval-2017 task 12: Clinical tempeval. In *Proceedings of SemEval-2017*. Association for Computational Linguistics, Vancouver, Canada, pages 563–570.

Michael Collins. 2002. Ranking algorithms for named-entity extraction: Boosting and the voted perceptron. In *Proceedings of ACL*. Association for Computational Linguistics, pages 489–496.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of machine learning research* 9(Aug):1871–1874.

Yoav Freund and Robert E Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine Learning* 37(3):277–296.

William A Gale and Geoffrey Sampson. 1995. Good-turing frequency estimation without tears*. *Journal of Quantitative Linguistics* 2(3):217–237.

Inc. Gurobi Optimization. 2015. Gurobi optimizer reference manual. http://www.gurobi.com.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proceedings of ACL*. Association for Computational Linguistics, volume 7, pages 264–271.

Artuur Leeuwenberg and Marie-Francine Moens. 2017. Structured learning for temporal relation extraction from clinical records. In *Proceedings of EACL*. Association for Computational Linguistics, Valencia, Spain.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *Proceedings of COLING–ACL*. Association for Computational Linguistics, pages 753–760.

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.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of NAACL-HLT*. Association for Computational Linguistics, pages 173–180.