LIMSI-COT at SemEval-2017 Task 12: Neural Architecture for Temporal Information Extraction from Clinical Narratives

Julien Tourille[1,2,3], **Olivier Ferret**[4], **Xavier Tannier**[1,2,3], **Aurélie Névéol**[1,3]

[1] LIMSI, CNRS
[2] Univ. Paris-Sud
[3] Université Paris-Saclay
[4] CEA, LIST, Gif-sur-Yvette, F-91191 France.
firstname.lastname@limsi.fr, olivier.ferret@cea.fr

Abstract

In this paper we present our participation to SemEval 2017 Task 12. We used a neural network based approach for entity and temporal relation extraction, and experimented with two domain adaptation strategies. We achieved competitive performance for both tasks.

1 Introduction

SemEval 2017 Task 12 offers 6 subtasks addressing medical event recognition and temporal reasoning in the clinical domain using the THYME corpus (Styler IV et al., 2014). Similarly to the two previous editions of the challenge (Bethard et al., 2015, 2016), the first group of subtasks concerns medical event (`EVENT`) and temporal expression (`TIMEX3`) extraction from raw text. In a second group of subtasks, participants are challenged to extract containment (`CONTAINS`) relations between `EVENT` and/or `TIMEX3` as well as Document Creation Time (DCT) relations between `EVENT` entities and documents in which they are embedded. The novelty of the 2017 edition lies in the difference of domains between train and test corpora. More details about the task and the definition of each subtask can be found in Bethard et al. (2017).

The task has been offered by SemEval over the past two years. Concerning the first group of subtasks, different approaches have been implemented by the participants including Conditional Random Fields (CRF) (AAl Abdulsalam et al., 2016; Caselli and Morante, 2016; Chikka, 2016; Cohan et al., 2016; Grouin and Moriceau, 2016; Hansart et al., 2016) and deep learning models (Fries, 2016; Chikka, 2016; Li and Huang, 2016). Similarly, CRF and neural networks models have been used for the second group of subtasks (AAl Abdulsalam et al., 2016; Cohan et al., 2016; Lee et al., 2016). Other approaches include Support Vector Machines (SVM) (AAl Abdulsalam et al., 2016; Tourille et al., 2016).

2 Methodology

The `EVENT` and `TIMEX3` entity extraction subtasks can be seen as two sequence labeling problems where each token of a given sentence is assigned a label. Entities can spread over several tokens and therefore, we used the IOB format (Inside, Outside, Beginning) for label representation. Each token can be at the beginning of an entity (B), inside an entity (I) or outside (O). `EVENT` entities are characterized by a *type* attribute that we used in our IOB scheme resulting in 7 possible labels. Similarly, `TIMEX3` entities are characterized by a *class* attribute that we used in our IOB scheme resulting in 13 possible labels.

The container relation extraction task can be cast as a 3-class classification problem. For each combination E1 – E2 of `EVENT` and/or `TIMEX3` from left to right, three cases are possible:

- E1 temporally *contains* E2,
- E1 *is* temporally *contained* by E2,
- there is no relation between E1 and E2.

Intra- and inter-sentence relation detection can be seen as two different tasks with specific features. Intra-sentence relations can benefit from intra-sentential clues such as adverbs (e.g. *during*) or pronouns (e.g. *which*) which are not available at the inter-sentence level. Furthermore, past work on the topic seems to indicate that this differentiation improves overall performance (Tourille et al., 2016). We have adopted this approach by building two separate classifiers, one for intra-sentence relations and one for inter-sentence relations.

If we were to consider all combinations of entities within documents for inter-sentence relations, it would result in a very large training corpus with very few positive examples. In order to cope with this issue, we limit our experiments to inter-

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 597–602,
Vancouver, Canada, August 3 - 4, 2017.

sentence relations that do not span over more than three sentences. By doing so, we obtain a manageable training corpus size with less unbalanced classes while keeping a good coverage.

3 Corpus Preprocessing

We preprocessed the corpus using cTAKES 3.2.2 (Savova et al., 2010), an open-source natural language processing system for the extraction of information from electronic health records. We extracted sentence and token boundaries, as well as token types and semantic types of the entities that have a span overlap with a least one gold standard EVENT entity of the THYME corpus. This information was added to the set of gold standard attributes available for EVENT entities in the corpus.

We also preprocessed the corpus using HeidelTime 2.2.1 (Strötgen and Gertz, 2015), a multilingual domain-sensitive temporal tagger, and used the results to further extend our feature set.

4 Models

4.1 Entity Extraction

Our approach relies on Long Short-Term Memory Networks (LSTMs) (Hochreiter and Schmidhuber, 1997). The architecture of our model is presented in Figure 1. For a given sequence of tokens, represented as vectors, we compute representations of left and right contexts of the sequence at every token. These representations are computed using two LSTMs (forward and backward LSTM in figure 1). Then these representations are concatenated and linearly projected to a n-dimensional vector representing the number of categories. Finally, as Huang et al. (2015), we add a CRF layer to take into account the previous label during prediction. Following preliminary experiments, we built one specific classifier for each entity type (EVENT or TIMEX3).

4.2 Event Attribute and Document Creation Time Relation Extraction

We treated each EVENT attribute (*ContextualModality*, *Degree*, *Polarity*) extraction subtask as a supervised classification problem. We built a common architecture for all attributes based on a linear SVM. Concerning DCT relation extraction subtask, we used the same architecture. We trained a separate classifier for each of the four subtasks based on lexical, contextual and structural features extracted from the documents:

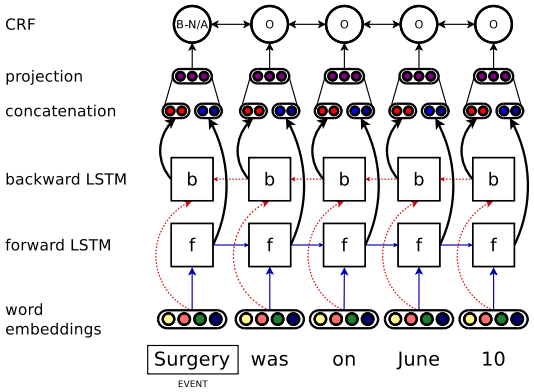

Figure 1: Neural model for EVENT extraction.

- EVENT type attribute,
- EVENT plain lexical form,
- EVENT position within the document,
- POS tags of the verbs within the right and left contexts of the considered entity,
- EVENT POS tag,
- type or class of the other entities that are present within the left and right contexts,
- token unigrams and bigrams within a window around the entity.

4.3 Temporal Relation Extraction

Similarly to our entity extraction approach, we built a system based on LSTMs for CONTAINS relation extraction. The architecture of our model is presented in Figure 2. For a given sequence of tokens between two entities (EVENT and/or TIMEX3), we compute a representation by scanning the sequence from left to right (forward LSTM in Figure 2). As LSTMs tend to be biased toward the most recent inputs, this model is biased toward the second entity of each pair processed by the network. To counteract this effect, we compute the reverse representation with an LSTM reading the sequence from right to left (backward LSTM in Figure 2). The two final states are then concatenated and linearly transformed into a 3-dimensional vector representing the number of categories (concatenation and projection in figure 2). Finally, a softmax function is applied.

4.4 Input Word Embeddings

Input vectors are built differently depending on the subtask. For the entity extraction subtask, vectors representing tokens are built by concatenating a character-based embedding and a word embedding. Whether we are dealing with EVENT or TIMEX3 entities, we add one embedding per

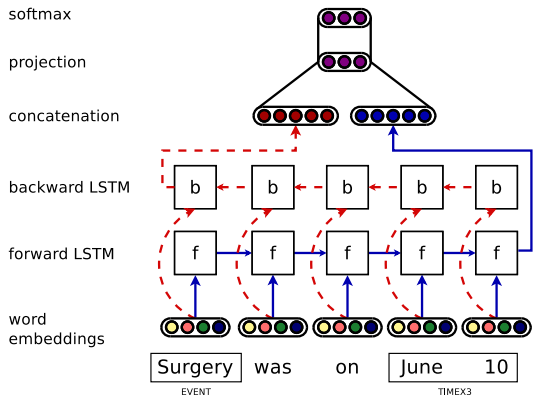

Figure 2: Neural architecture for CONTAINS relation extraction.

cTAKES attribute or one embedding representing the TIMEX3 *class* as detected by HeidelTime. Concerning the containment relation subtask, input vectors are built by concatenating a character-based embedding, a word embedding, one embedding per Gold Standard attribute and one embedding for the type of DCT relations (*before* ...).

An overview of the embedding computation is presented in Figure 3. Following Lample et al. (2016), the character-based representation is constructed with a Bi-LSTM[1]. First, a random embedding is generated for every character in the training corpus. Token characters are then processed with a forward and backward LSTM architecture similar to the one of our entity extraction model. The final character-based representation results from the concatenation of the forward and backward representations. Since medical terms often include prefixes and suffixes derived from ancient Greek and classical Latin (Namer and Zweigenbaum, 2004), we believe that both entity and containment relation extractions can particularly benefit from this character-based representation of tokens for terms that have not been seen during training or that don't have a pretrained word embedding.

We use pretrained word embeddings computed with *word2vec* (Mikolov et al., 2013)[2] on the Mimic 3 corpus (Johnson et al., 2016) and the colon cancer part of the THYME corpus. In order to account for unknown tokens during the test phase, we train a special embedding UNK by replacing randomly some singletons with the UNK embedding (probability of replacement = 0.5). In

[1]Embedding size = 8; hidden layer size = 25.

[2]Parameters used during computation: algorithm = CBOW; min-count = 4; vector size = 100; window = 8.

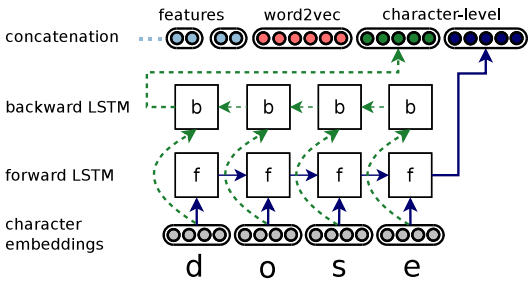

Figure 3: Model for character-based embeddings.

the inter-sentence relation classifier, we introduce a specific token for identifying sentence breaks. This token is composed of one distinctive character and it is associated to a specific word embedding. Similarly to the character embeddings, we randomly initialize one embedding per token attribute value, with an embedding size of 4. All these embeddings are then concatenated.

4.5 Network Training

We implemented the two neural networks models described in the previous section using TensorFlow 0.12 (Abadi et al., 2015). We trained our networks with mini-batch Stochastic Gradient Descent using Adam (Kingma and Ba, 2014)[3]. We use dropout training to avoid overfitting. We apply dropout on input embeddings with a rate of 0.5.

The optimization of hyperparameters for the attribute and DCT relation extraction subtasks was addressed by using a Tree-structured Parzen Estimator approach (Bergstra et al., 2011) and applied to the hyperparameter C of the linear SVM, the lookup window around entities and the percentile of features to keep. For the latter we used the ANOVA F-value as selection criterion.

5 Domain Adaptation Strategies

We implemented two strategies for domain adaptation during the first phase. In the first strategy, we blocked further training of the pretrained word embeddings during network training. Since a large number of medical events mentioned in the test set are not seen during training, we believe that our system should rely on untuned word embeddings to make its prediction.

In the second strategy we randomly replaced tokens that composed EVENT entities by the *unknown* token[4]. Given the fact that our word em-

[3]Learning rate = 0.001; hidden layer sizes = 256.

[4]Replacement probability = 0.2.

	Phase 1						Phase 2					
	STATIC			REPLACE			ALL			30-30		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
EVENT Span	.622	.843	**.716**	.606	.841	.705	.691	.854	**.764**	.660	.865	.749
EVENT Modality	.553	.749	**.636**	.537	.745	.624	.628	.775	**.694**	.598	.784	.679
EVENT Degree	.616	.834	**.708**	.600	.831	.697	.682	.843	**.754**	.652	.854	.739
EVENT Polarity	.603	.816	**.693**	.588	.815	.683	.676	.835	**.747**	.644	.844	.731
EVENT Type	.608	.823	**.699**	.592	.821	.688	.675	.834	**.746**	.641	.841	.728
EVENT All attributes	.374	.507	**.431**	.365	.507	.425	.468	.578	**.517**	.440	.577	.500
TIMEX3 Span	.421	.660	.514	.421	.660	.514	.510	.671	**.579**	.452	.621	.523
TIMEX3 Class	.401	.630	.490	.401	.630	.490	.487	.641	**.553**	.430	.591	.498
DCT Relation	.443	.599	**.509**	.436	.604	.506	.535	.661	**.591**	.511	.670	.580
CONTAINS	.280	.396	**.328**	.264	.408	.320	.244	.438	**.316**	.211	.422	.282

Table 1: Results obtained by our system across our four runs. We report Precision (P), Recall (R) and F1-measure (F1). The best F1 performance in each phase is bolded.

beddings are pretrained on the Mimic 3 corpus and on the colon cancer part of the THYME corpus, a number of tokens (and therefore EVENTs) of the test part of the corpus may not have a specific word embedding. By replacing randomly EVENT token, we force our networks to look at other contextual clues within the sentence. Both strategies were applied on EVENT entity and CONTAINS relation extraction subtasks.

Phase 2 was addressed by implementing two strategies. In the first one, we mixed the 30 texts about brain cancer to the 591 texts about colon cancer. In the second one, we randomly chose 30 texts related to colon cancer and combined them to the 30 texts about brain cancer, resulting in a balanced training corpus. Both strategies were applied on EVENT, TIMEX3 and CONTAINS extraction subtasks.

6 Results and Discussion

Results for our four runs are presented in Table 1. The two strategies implemented for Phase 1 yield similar results (0.01 difference in F1-measure at most), with only a very slight advantage for the strategy blocking further training of the word embeddings (STATIC strategy in the table). In Phase 2, the two strategies also yield close results (0.04 difference in F1-measure) for the EVENT entity extraction and temporal relation subtasks. However, the strategy consisting in taking all available annotations (ALL strategy in the table) outperforms slightly the training on a balanced corpus, especially for the extraction of CONTAINS relations. The same strategy seems to perform much better for the TIMEX3 entity extraction subtask where the gap in F1-measure reaches 0.06. This superiority agrees the general observation that the size of the training corpus has often a greater impact on results than its strict matching with the target domain. Overall, in both phases and for all strategies, results are competitive for entity and temporal relation extraction.

The performance obtained by our system relies in part on corpus tayloring. Some sections of the test corpus related to *medication* and *diet* are not to be annotated according to the annotation guidelines. However, these sections are not formally delimited within the documents. To avoid annotating them during test time, we developed a semi-automatic approach for detecting these sections and put them aside.

Other aspects linked to the corpus limit the performance. Some sections should not be annotated as they are duplicate of other sections found in the corpus as a whole. However, we have no information on how to formally identify these sections. Furthermore, a number of temporal expressions are annotated as SECTIONTIME or DOCTIME entities. Detecting TIMEX3 entities instead decreases the precision of our model.

In future work, we plan to explore additional strategies. For instance, adding a feature predicting whether a given EVENT entity is a container or not has proved useful in previous work (Tourille et al., 2016), but was not implemented in our system due to time constraints.

Acknowledgements This work was supported by Labex Digicosme, operated by the Foundation for Scientific Cooperation (FSC) Paris-Saclay, under grant CÔT.

References

Abdulrahman AAl Abdulsalam, Sumithra Velupillai, and Stephane Meystre. 2016. UtahBMI at SemEval-2016 Task 12: Extracting Temporal Information from Clinical Text. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 1256–1262.

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org.

James S. Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for Hyper-Parameter Optimization. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, Curran Associates, Inc., pages 2546–2554.

Steven Bethard, Leon Derczynski, Guergana Savova, 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, Denver, USA, pages 806–814.

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)*. Association for Computational Linguistics, San Diego, California, pages 1052–1062.

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. Semeval-2017 task 12: Clinical tempeval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada.

Tommaso Caselli and Roser Morante. 2016. VUA-CLTL at SemEval 2016 Task 12: A CRF Pipeline to Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 1241–1247.

Veera Raghavendra Chikka. 2016. CDE-IIITH at SemEval-2016 Task 12: Extraction of Temporal Information from Clinical documents using Machine Learning techniques. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 1237–1240.

Arman Cohan, Kevin Meurer, and Nazli Goharian. 2016. Guir at semeval-2016 task 12: Temporal information processing for clinical narratives. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 1248–1255.

Jason Fries. 2016. Brundlefly at semeval-2016 task 12: Recurrent neural networks vs. joint inference for clinical temporal information extraction. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 1274–1279.

Cyril Grouin and Véronique Moriceau. 2016. LIMSI at SemEval-2016 Task 12: machine-learning and temporal information to identify clinical events and time expressions. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 1225–1230.

Charlotte Hansart, Damien De Meyere, Patrick Watrin, André Bittar, and Cédrick Fairon. 2016. CENTAL at SemEval-2016 Task 12: a linguistically fed CRF model for medical and temporal information extraction. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 1286–1291.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint* arXiv:1508.01991.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo A. Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, San Diego, California, pages 260–270.

Hee-Jin Lee, Hua Xu, Jingqi Wang, Yaoyun Zhang, Sungrim Moon, Jun Xu, and Yonghui Wu. 2016. UTHealth at SemEval-2016 Task 12: an End-to-End System for Temporal Information Extraction from Clinical Notes. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 1292–1297.

Peng Li and Heng Huang. 2016. Uta dlnlp at semeval-2016 task 12: Deep learning based natural language processing system for clinical information identification from clinical notes and pathology reports. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 1268–1273.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781. http://arxiv.org/abs/1301.3781.

Fiammetta Namer and Pierre Zweigenbaum. 2004. Acquiring meaning for French medical terminology: contribution of morphosemantics. In *Proceedings of the 10th World Congress on Medical Informatics (MEDINFO)*. pages 535–539.

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.

Jannik Strötgen and Michael Gertz. 2015. A Baseline Temporal Tagger for all Languages. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 541–547.

William Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics* 2:143–154.

Julien Tourille, Olivier Ferret, Aurélie Névéol, and Xavier Tannier. 2016. LIMSI-COT at SemEval-2016 Task 12: Temporal relation identification using a pipeline of classifiers. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 1136–1142.