# LIMSI at SemEval-2016 Task 12: machine-learning and temporal information to identify clinical events and time expressions

**Cyril Grouin**
LIMSI, CNRS, Université Paris-Saclay
Bât 508, rue John von Neumann, Campus Universitaire, F-91405 Orsay
cyril.grouin@limsi.fr

**Véronique Moriceau**
LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay
Bât 508, rue John von Neumann, Campus Universitaire, F-91405 Orsay
veronique.moriceau@limsi.fr

## Abstract

Our experiments rely on a combination of machine-learning (CRF) and rule-based (HeidelTime) systems. First, a CRF system identifies both EVENTS and TIMEX3, along with polarity values for EVENT and types of TIMEX. Second, the HeidelTime tool identifies DOCTIME and TIMEX3 elements, and computes DocTimeRel for each EVENT identified by the CRF. Third, another CRF system computes DocTimeRel for each previously identified EVENT, based on DocTimeRel computed by HeidelTime. In the first submission, all EVENTS and TIMEX3 are identified through one general CRF model while in the second submission, we combined two CRF models (one for both EVENT and TIMEX3, and one only for TIMEX3) and we applied post-processing rules on the outputs.

## 1 Introduction

In this paper, we present the methods we used while participating in the 2016 Clinical TempEval task as part of the SemEval-2016 challenge. A few recent NLP challenges focused on temporal expressions within clinical records, such as the 2014 i2b2/UTHealth[1] challenge (Stubbs et al., 2015) or the second task from the 2014 ShARe/CLEF eHealth[2] evaluation lab (Mowery et al., 2014).

## 2 Task description

### 2.1 Presentation

The 2016 Clinical TempEval track[3] proposed six tasks (Bethard et al., 2016). We participated in the first five tasks which concern the identification of: $(i)$ spans of time expressions (TS task), $(ii)$ spans of event expressions (ES task), $(iii)$ the attribute of time expressions (TA task), $(iv)$ attributes of event expressions (EA task), and $(v)$ the relation between each event and the document creation time (DR task). We did not participate in the narrative container relation task (CR).

In both TS and ES tasks, spans of time and event expressions are represented using start and end offsets of characters. In the TA task, time expressions are related to the attribute "class" which specifies the type of time expressions among six possible values: *date, duration, prepostexp, quantifier, set, time*. In the EA task, event expressions are related to four attributes: "polarity" (either *positive* or *negative*), "modality" (*actual, hedged, hypothetical* or *generic*), "degree" (*most, little* or *N/A*) and "type" (*aspectual, evidential* or *N/A*).

Two phases were proposed. In the first phase, all tasks were proposed, based on raw texts. We thus participated to TS ES TA EA and DR tasks. In the second phase, reference annotations were given for tasks TS ES TA and EA, and participants were expected to identify the relations from DR and CR tasks. We also participated in this DR task.

[1] https://www.i2b2.org/NLP/TemporalRelations/

[2] http://clefehealth2014.dcu.ie/task-2

[3] http://alt.qcri.org/semeval2016/task12/

*Proceedings of SemEval-2016*, pages 1225–1230,
San Diego, California, June 16-17, 2016. ©2016 Association for Computational Linguistics

### 2.2 Corpora

The given training corpus was divided into two subcorpora called *train* and *dev*. Table 1 shows the distribution of each category and its attributes in this corpus. The most frequent ones are in bold font.

| | **Attributes** | **Train** | **Dev** | **TOTAL** |
|---|---|---|---|---|
| TIMEX | **Class** | | | |
| | Date | 2588 | 1422 | **4010** |
| | Duration | 434 | 200 | 634 |
| | PrePostExp | 313 | 172 | 485 |
| | Set | 218 | 116 | 334 |
| | Quantifier | 162 | 109 | 271 |
| | Time | 118 | 59 | 177 |
| EVENT | **Type** | | | |
| | N/A | 36185 | 19414 | **55599** |
| | Evidential | 2206 | 1314 | 3520 |
| | Aspectual | 546 | 246 | 792 |
| | **Polarity** | | | |
| | POS | 34832 | 18795 | **53627** |
| | NEG | 4105 | 2179 | 6284 |
| | **Degree** | | | |
| | N/A | 38698 | 20864 | **59562** |
| | Little | 143 | 65 | 208 |
| | Most | 96 | 45 | 141 |
| | **Modality** | | | |
| | Actual | 35781 | 22647 | **58428** |
| | Hypothetical | 1656 | 829 | 2485 |
| | Hedged | 889 | 443 | 1332 |
| | Generic | 611 | 299 | 910 |
| | **DocTimeRel** | | | |
| | Overlap | 18297 | 9812 | **28109** |
| | Before | 14291 | 7896 | 22187 |
| | After | 4189 | 2138 | 6327 |
| | Before/overlap | 2160 | 1128 | 3288 |

**Table 1:** Statistics on the training data.

Corpora are composed of files of two types: "clinic" and "path" files. According to the organizers, annotations are of better quality for "clinic" files than for "path" files. Moreover, adjudication of annotations has been made only for "clinic" files.

## 3 Methods

Our methods mainly rely on machine-learning. We used the Wapiti (Lavergne et al., 2010) toolkit, based on the linear-chain CRFs framework (Lafferty et al., 2001). We considered this challenge as a classification task, where we have to classify each token from a file into a TIMEX or an EVENT category.

Due to the unbalanced distribution of attribute values in the training corpus (see Table 1), we decided not to automatically process three EVENT attributes ("modality", "degree", and "type"). For those attributes, we used the most frequent value as a default value: *Actual* for "modality" (89.5%), *N/A* for both "degree" (99.4%) and "type" (92.8%) attributes. As a consequence, we only processed the "polarity" and "doctimerel" EVENT attributes using our CRF systems.

### 3.1 Tasks TS, ES, TA, EA

**CRF system** We merged all tasks of spans and attributes identification into a single task. This process consists in identifying the main category and the related attribute value in one step (e.g., EVENT-POS and EVENT-NEG to identify *positive* and *negative* EVENT expressions).

Since annotations are of better quality in "clinic" files, we compared the results we achieved (ES and TS tasks) on the whole development set (i.e., both "clinic" and "path" files) whether we trained our CRF model on both "clinic" and "path" files from the training set or on the "clinic" files from this training set only. Table 2 presents those results. We observed the CRF model only trained on the "clinic" files outperforms results globally and for each category. As a consequence, we decided to train our CRF models on the "clinic" files only.

| Training set | Category | P | R | F |
|---|---|---|---|---|
| ClinPath<br>(293 files) | EVENT | **.877** | .710 | .785 |
| | TIMEX | .801 | .517 | .629 |
| | Overall | **.872** | .693 | .773 |
| Clin<br>(195 files) | EVENT | .845 | **.869** | **.857** |
| | TIMEX | **.810** | **.551** | **.656** |
| | Overall | .843 | **.842** | **.843** |

**Table 2:** Results on the development set whether the CRF model was trained on both "clinic" and "path" files (ClinPath) from the training set or "clinic" files only (Clin) (P=Precision, R=Recall, F=F-measure). Black font highlights the best results

We thus created three CRF models:

- a first model to identify both spans of EVENT expressions and the associated value of the attribute of polarity: model EVENT/Polarity;
- a second model to identify both spans of

TIMEX expressions and the associated value of the “class” attribute: model TIMEX/Class;

- and a last global model to identify both EVENT and TIMEX expressions (i.e., spans) with the associated values of “polarity” and “class” attributes: model EVENT/Polarity TIMEX/Class.

The following features were used to produce all CRF models: $(i)$ the token itself; $(ii)$ token length, typographic case of the token, presence of punctuation marks in the token, and presence of digits in the token; $(iii)$ part-of-speech tag of the token, provided by the Tree Tagger POS tagger (Schmid, 1994); $(iv)$ cluster ID of each token through an automatic unsupervised clustering of all tokens from the training corpus into 120 clusters,[4] using the algorithm designed by Brown et al. (1992) and implemented by Liang (2005).

**Rule-based post-processing** In order to correct predictions made by the CRF system, we designed a basic post-processing based on rules. This post-processing consists in identifying additional EVENT expressions from a list of 69 most frequent EVENT expressions we collected from the training set. For all EVENT expressions found in this list, we set the “Polarity” value to *positive* as a default value.

### 3.2 Task DR

**HeidelTime tool** HeidelTime (Strötgen and Gertz, 2010; Strötgen and Gertz, 2013) allows to extract and normalize temporal expressions in texts according to the TIMEX3 standard. For the normalization of relative temporal expressions, the document creation time (DCT) is used. In this task, HeidelTime considers the expression indicated as *start date* in the text files as the DCT. Once all TIMEX are normalized, we assigned to each of them one of the four temporal relations: *before, after, overlap, before/overlap* w.r.t the DCT.

**CRF system** We considered this relation task as a classification task, where we have to classify each EVENT into a relational class. In order to perform this classification, we gave the CRF additional features, namely the temporal features previously computed by the HeidelTime tool on each TIMEX.

## 4 Design of experiments

Figure 1 presents the design of experiments we followed for the official submissions for the first phase. The grey boxes represent the distinct CRF models we used and the type of expressions processed by each model.

### 4.1 Task TS, TA, ES, EA (phase 1)

We considered two configurations, based on the results we achieved on the development corpus:

- application of the global model to identify all elements (EVENT/Polarity and TIMEX/Class). In our experiments, we noticed this model allows us to achieve higher values for both precision and F-measure. This constitutes our first submission (run #1.1);
- application of both global model (i.e., EVENT/Polarity and TIMEX/Class) and TIMEX/Class specific model to identify all elements. We merged outputs from both models, giving priority to predictions from the specific model for the TIMEX/Class category, and applied our post-processing. This configuration allows us to obtain higher values for recall. This constitutes our second submission (run #1.2).

### 4.2 Task DR

#### 4.2.1 Task DR (phase 1)

For the first phase, we designed a new CRF model to predict the value of the “DocTimeRel” attribute associated with each EVENT we previously identified. In order to improve the quality of predictions made by our CRF system, we added as features the temporal relations computed by HeidelTime (see Section 3.2): the relations computed by HeidelTime are mapped to TIMEX expressions detected by the CRF models used in phase 1. When the CRF model detects a TIMEX expression that is not extracted by HeidelTime, the default value is used (*overlap*).

[4] We used five distinct versions of each cluster ID: the original cluster ID (e.g., “01011”) and four generalization of each ID, removing the last digit to produce a more generic version from the previous one for each iteration (“0101”, “010”, “01” and “0”). We gave the CRF all those versions.

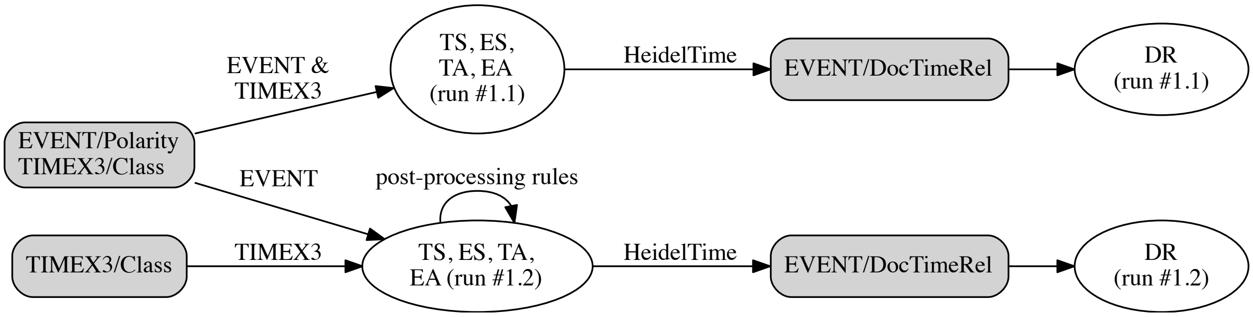

**Figure 1:** Design of experiments for phase 1. Grey boxes represent our distinct CRF models

We applied this model on outputs from runs #1.1 and #1.2. In this experiment, we tested how performs our DR CRF model depending on the quality of the outputs from the previous TS/ES/TA/EA tasks. Additionally, we considered the relations computed by HeidelTime and associated with each TIMEX expression would be useful.

### 4.2.2 Task DR (phase 2)

For the second phase, since gold standard annotations for TIMEX and EVENT expressions were given by the organizers, we trained two CRF models based on two different inputs:

- run #2.1: temporal relations computed by HeidelTime are mapped to the TIMEX entities from the gold standard annotations. If a TIMEX is not extracted by HeidelTime, the default value is used (*overlap*);
- run #2.2: only TIMEX and temporal relations extracted by HeidelTime are used.

## 5 Results and discussion

### 5.1 Official results (test set)

In this section, we present the official results we achieved on the test set. Outputs are evaluated through classical metrics used in information retrieval tasks: precision (positive predictive value), recall (true positive rate), and F-measure (weighted harmonic mean of precision and recall). In order to make the comparison between the results we obtained among our two submissions easier, the bold font pinpoints our best results.

### 5.1.1 Phase 1

Table 3 presents the official results we achieved on the test set for the tasks TS, ES, TA, EA and DR based on our first submission (run #1.1) from phase 1.

| Task | Category | P | R | F |
|---|---|---|---|---|
| TS | TIMEX | **.840** | .510 | .635 |
| ES | EVENT | **.885** | .808 | **.845** |
| TA | TIMEX/Class | **.815** | .495 | .616 |
| EA | EVENT/Degree | **.880** | .805 | **.841** |
| | EVENT/Modality | **.811** | .742 | **.775** |
| | EVENT/Polarity | **.867** | .792 | **.828** |
| | EVENT/Type | **.825** | .754 | **.788** |
| DR | EVENT/DocTimeRel | **.635** | .580 | **.607** |

**Table 3:** Official results on the test set (run #1.1) for tasks TS, ES, TA, EA and DR (P = Precision, R = Recall, F = F-measure). Bold font pinpoints best results

Table 4 presents the official results we achieved on the test set for the tasks TS, ES, TA, EA and DR based on our second submission (run #1.2) from phase 1.

| Task | Category | P | R | F |
|---|---|---|---|---|
| TS | TIMEX | .830 | **.518** | **.638** |
| ES | EVENT | .869 | **.816** | .842 |
| TA | TIMEX/Class | .804 | **.503** | **.619** |
| EA | EVENT/Degree | .865 | **.812** | .838 |
| | EVENT/Modality | .798 | **.749** | .772 |
| | EVENT/Polarity | .851 | **.799** | .824 |
| | EVENT/Type | .811 | **.761** | .785 |
| DR | EVENT/DocTimeRel | .624 | **.585** | .604 |

**Table 4:** Official results on the test set (run #1.2) for tasks TS, ES, TA, EA and DR (P = Precision, R = Recall, F = F-measure). Bold font pinpoints best results

We achieved our best results in our first submission, which is based on a global CRF model for tasks TS, ES, TA and EA, and a second CRF global model for task DR. As expected in our experimental setup, our first submission allows us to maximize the precision values, and generally obtains the higher F-measure values, while the second submission maximizes the recall values for all category and attribute.

In comparison with submissions from other participants from this first phase, we succeed to obtain better results than the median F-measure value for tasks TS (median=.637 vs. F=.638 in run #1.2) and ES (median=.830 vs. F=.845 and .842 respectively in runs #1.1 and #1.2). On TA, EA and DR tasks, we obtained lower results w.r.t. the median values. Our system only succeeds to obtain better results than the organizers' baseline on tasks TS (baseline F=.551) and DR (baseline=.604 vs. F=.607 in run #1.1).

### 5.1.2 Phase 2

Table 5 presents the official results we achieved on the test set for the task DR based on our first submission (run #2.1) from phase 2.

| **Task** | **Category** | **P** | **R** | **F** |
|---|---|---|---|---|
| DR | EVENT/DocTimeRel | .687 | .687 | .687 |

**Table 5:** Official results on the test set (run #2.1) for task DR (P = Precision, R = Recall, F = F-measure)

Table 6 presents the official results we achieved on the test set for the task DR based on our second submission (run #2.2) from phase 2.

| **Task** | **Category** | **P** | **R** | **F** |
|---|---|---|---|---|
| DR | EVENT/DocTimeRel | .679 | .679 | .679 |

**Table 6:** Official results on the test set (run #2.1) for task DR (P = Precision, R = Recall, F = F-measure)

On both submissions from phase 2, we did not succeed to obtain better results than the median value (F=.724). Nevertheless, our system obtained better results than the organizers' baseline (F=.675 vs. F=.687 in run #2.1 and .679 in run #2.2). This observation is consistent with the results we achieved on the DR task in phase 1.

## 5.2 Error analysis

**Lack of robustness** Since we trained our CRF models on "clinic" files only (see section 3.1), most of false negatives concern EVENT expressions that are only used in "path" files (e.g., *Description, nodes, orientation*, etc.). As the CRF failed to identify those expressions, we can consider that our model failed to generalize the properties of EVENT expressions, despite other features, and experienced difficulties to process unknown elements.

Similarly, due to the differences of structure between "clinic" and "path" files, most of false positives concern predictions made by our CRF system on "path" files exclusively (e.g., our system considered *Hematoxylin* to be annotated as an EVENT in all "path" files). This observation highlights the lack of robustness of our CRF systems on files being of different type.

Table 7 presents the results we achieved on the test set for each "clinic" and "path" files sub-set. Bold font highlights the best results. As expected, results are better on the "clinic" files sub-set than on the "path" files sub-set (on average 4 points more on the EVENT category for each metric).

| Test sub-set | Category | P | R | F |
|---|---|---|---|---|
| Clinic<br>(102 files) | EVENT | **.871** | **.891** | **.881** |
| | TIMEX | **.811** | **.663** | **.730** |
| | Overall | **.866** | **.867** | **.866** |
| Path<br>(51 files) | EVENT | .828 | .853 | .840 |
| | TIMEX | .000 | .000 | .000 |
| | Overall | .828 | .848 | .838 |

**Table 7:** Results on the test set for each "clinic" and "path" files sub-set (P=Precision, R=Recall, F=F-measure)

Additionally, we observed that our CRF model failed to identify any TIMEX expressions on the "path" files from the test set.

**TIMEX extraction with HeidelTime** Evaluation of HeidelTime on TIMEX extraction of the test set gives the following results: P=.479, R=.586 and F=.527. The best scores are obtained for type *date* (P=.687, R=.677 and F=.682) whereas scores are very low for *quantifier* (F=.043), *set* (F=.065) and *time* (F=.083). This can be explained by the fact that some TIMEX annotations in the gold data are not compliant with the TimeML standard as used in HeidelTime. For example, the expression *12-MAY-2001 21:11* is annotated as a date (*12-MAY-2001*) and a time (*21:11*) in the gold data whereas the whole expression should be annotated as a time according to the TimeML standard. These low scores also explain why only 75% of TIMEX expressions in the gold data are mapped with a temporal relation computed by HeidelTime in run #2.1.

## 6 Conclusion

In this paper, we presented the methods we used while participating in the 2016 Clinical TempEval task as part of the SemEval-2016 challenge. We considered each task from the challenge as a classification task using machine-learning approaches. Our CRF systems allow us to predict both the offsets of each expression (TS and ES tasks) and the attribute of expressions (TA and EA tasks) as well as the temporal relations (DR task). Due to an unbalanced distribution of some EVENT attributes (namely "degree", "modality" and "type"), we gave the most frequent used value in the training corpus as a default value.

We designed two experiments. The first one is based on a single CRF model which identifies all expressions and attributes at the same time. The second one is based on a merge of two CRF models (one for EVENT, the other for TIMEX) and post-processing rules to identify new EVENT expressions.

We achieved our best results through the first experiment (higher precision values for all tasks and the best F-measure values for all EVENT related tasks, F=.845 on ES) while the second experiment allows us to obtain higher recall values for all tasks and the best F-measure values for all TIMEX related tasks (F=.638 on TS). Our systems succeed to identify more correctly EVENT expressions than TIMEX expressions.

Our system achieved better results on the "clinic" files sub-set rather than on the "path" files sub-set. Future work is thus needed, first to improve the generalization of features used (in order to predict unknown expressions based on features), and second to ensure the robustness of our system (in order to process different files, namely "path" files).

## References

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.

Peter F Brown, Vincent J Della Pietra, Peter V de Souza, Jenifer C Lai, and Robert L Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–79.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proc of ICML*, pages 282–9, Williamstown, MA.

Thomas Lavergne, Olivier Cappé, and François Yvon. 2010. Practical very large scale CRFs. In *Proc of ACL*, pages 504–13, Uppsala, Sweden, July.

Percy Liang. 2005. Semi-supervised learning for natural language. Master's thesis, Massachusetts Institute of Technology.

Danielle L Mowery, Sumithra Velupillai, Brett R South, Lee Christiensen, David Martinez, Liadh Kelly, Lorraine Goeuriot, Noémie Elhadad, Sameer Pradhan, Guergana K Savova, and Wendy W Chapman. 2014. Task 2: ShARe/CLEF eHealth evaluation lab 2014. In *CLEF2014 Working Notes*, volume 1180, Sheffield, UK.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proc of International Conference on New Methods in Language*.

Jannik Strötgen and Michael Gertz. 2010. Heideltime: high quality rule-based extraction and normalization of temporal expressions. In *Proc of SemEval*.

Jannik Strötgen and Michael Gertz. 2013. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298.

Amber Stubbs, Christopher Kotfila, Hua Xu, and Özlem Uzuner. 2015. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task track 2. *J Biomed Inform*, 58:S67–S77.