CENTAL at SemEval-2016 Task 12: A linguistically fed CRF model for medical and temporal information extraction

Charlotte Hansart Damien De Meyere Patrick Watrin
André Bittar Cédrick Fairon

CENTAL
Université catholique de Louvain
Place Blaise Pascal 1, 1348 Louvain-la-Neuve, Belgium

Abstract

In this paper, we describe the system developed for our participation in the Clinical TempEval task of SemEval 2016 (task 12). Our team focused on the subtasks of span and attribute identification from raw text and proposed a system that integrates both statistical and linguistic approaches. Our system is based on Conditional Random Fields with high-precision linguistic features.

1 Introduction

Extracting and linking temporal and medical information from medical documents is a highly useful task for many clinical applications and plays an important role in health care assessment and patient safety (Sun et al., 2013). Since 2007, the SemEval competition includes a temporal information extraction task which is now transposed onto the medical domain.

This year, the twelfth task included six subtasks that are described in (Bethard et al., 2016). Our team focused on the first four of these (Phase 1 submission), i.e. identification of entity spans and attribute values from raw text[1].

We present an approach combining linguistic rules and machine learning methods. The tools presented here were initially developed for French to extract information (temporal expressions, Named Entities, etc.) from newspaper corpora, a text genre which is quite different from medical documents. This challenge was thus a good opportunity for us to evaluate the scalability of our tools for both a language other than French and a new text genre.

[1]For the event attributes, only the "polarity" attribute is handled by our system.

2 Data

We received three different sets of medical documents: two for the development of our system, and one for testing purposes. The documents are clinical notes and pathology reports. The *Train* dataset comprises 297 clinical reports. The *Dev* dataset, used for testing our developments, contains 150 reports. Finally, the *Test* dataset, available only for the evaluation of our model, includes 153 documents.

All the documents are manually de-identified by the Mayo Clinic and annotated according to an extension of the ISO-TimeML standard (Pustejovsky et al., 2010): the THYME Annotation Guidelines, developed for the THYME project (Styler et al., 2014).

3 Methodology

The originality of our approach comes from the simultaneous consideration of both terminological and linguistic resources which feed a statistical model based on the Conditional Random Fields paradigm (Lafferty et al., 2001).

In our opinion, this strategy allowed us to take full advantage of the precision inherent to symbolic approaches while enabling us to benefit from the flexibility of supervised statistical modeling methods. Such a hybrid methodology has proven useful in previous research dealing with part-of-speech tagging (Constant and Sigogne, 2011).

Proceedings of SemEval-2016, pages 1177–1182,
San Diego, California, June 16-17, 2016.

3.1 Statistical analysis

Models such as Conditional Random Fields (CRF) are able to learn quickly and effectively from a large amount of observed data represented by features that are identified during the preprocessing step. Such systems typically use a set of language-independent and generic features, such as the prefix of the token or the characters that compose each token (letters, digits, hyphens, etc.).

Such a model has many advantages, the most important being its sequentiality, i.e. the fact that the CRF takes into account the context of an observation. But this approach also has its limitations. First, it is language- and domain-independent. Consequently, some generic features can be irrelevant for the concerned language or domain. Likewise, a standard CRF learns from its training corpus. Consequently, its knowledge is proportional to the variety of the corpus: if a word, a pattern, a structure is not in the corpus, the model will not recognise it and this might generate errors. In the case of textual data, this may also result in limited (lexical) coverage over unseen data.

To address the weaknesses of CRF learning, we used the CRF presented in (Watrin et al., 2014), which allows us to leverage external linguistic knowledge and resources, that will be presented in the next section. We looked at the span and attribute identification task as a sequence labelling task: a single model tags each token into the documents as being or not a relevant entity and the assigned tags capture the spans and the type of the recognised entities. The possible entity tags follow the pattern $\langle TYPE \rangle + \langle ATTRIBUTE \rangle$, so our tagset contains tags such as TIMEX3+DATE, TIMEX3+DURATION or EVENT+NEG. The O tag is assigned to tokens outside of any entity. In short, the sequence labelling tagging allows us to identify in a single run the boundaries, the type and the attribute of an entity.

The CRF model is developed with the CRFsuite package (Okazaki, 2007). Regarding its parameterisation, we used the default parameters proposed by the library (graphical model: first-order Markov CRF with dyad features; training algorithm: L-BFGS; see the library documentation for more information). The set of features remains unchanged against our original model (Watrin et al., 2014) and is reproduced in Table 1 and 2. As we will show in section 3.2, we tuned the *containsFeature(x)* feature according to the particularities of medical language.

Feature	explanation
Lexical item	token to be labeled
lowercase	token in lowercase
hasHyphen	does the token contain hyphen?
hasDigit	does the token contain digits?
allUpperCase	is the token uppercase only?
shape	token in a Xxx form
prefix*(n)*	*n* first letters of the token
suffix*(n)*	*n* last letters of the token

Table 1: Language-independent features

Feature	explanation
pos	token part-of-speech tag
containsFeature*(x)*	does the token belong to the semantic class *x* ?
sac	semantic class ambiguity (all possible classes for the token)

Table 2: Lexical features

3.2 Language resources

We combine two kinds of language-dependent resources: terminological and linguistic resources.

Terminological resources The language used in medical health records is characterized by a specific terminology. As this challenge focuses on English texts, we were able to reuse the numerous existing terminologies that are distributed throughout the *Unified Medical Language System* (UMLS). The UMLS brings together more than 150 vocabularies covering various aspects of healthcare (ICD-9-CM, ICD-10-CM, SNOMED-CT, MeSH, etc.) which were compiled into a dictionary of nearly 6 million terms. These entries were automatically classified into five generic categories, i.e. *procedures*, *diagnoses*, *anatomical terms*, *organisms* and *other* by simplifying the categories system of the UMLS. These categories were used as additional features for learning. As a word can be ambiguous, each category is considered as a separate feature in the CRF.

The training corpus being quite limited in size, these resources are necessary to overcome the problem of data sparseness as far as the medical terminology is concerned. As the UMLS already has an extensive coverage, we limited ourselves to importing its terminological components without any further linguistic processing, i.e. splitting into smaller units, stemming, etc. However, it has been shown that terminological components are only partially representative of the medical language in use (De Meyere et al., 2015). For example, terms

in the UMLS can be affected by syntagmatic and paradigmatic variation to varying degrees (e.g. *incomplete bladder emptying = the emptying function of the bladder is incomplete*) or may be too precise or complex to actually be used in electronic health records (e.g. *Histologically confirmed intracranial glioblastoma multiforme (GBM) or gliosarcoma*). The use of approximate string matching, generation of lexical variants and cleaning up of the dictionary would certainly improve the performance for this step and have a beneficial influence on the subsequent processes. We will discuss some points of further improvement in section 4.

Linguistic resources Besides terminological resources, we also developed linguistic rules for the detection of negated sequences (e.g. *Patient does not show any signs of dehydration*) and temporal expressions (e.g. *In 2006 and 2009*). These two types of information can be expressed in a large variety of ways and may be split into more than one part. We therefore formalized our rules into local grammars as implemented by the text processing framework Unitex (Paumier, 2003).

Concretely, these graphs model generic patterns that are manually constructed to extract relevant sequences. Figure 1 represents an example of graphs for the extraction of time points (e.g. *in 1998 and 2001*, *three years ago*, etc.) and duration (e.g. *during the last few weeks*). Grey boxes represent subgraphs that model days of the month, month names (and possible abbreviations), years, etc. Such grammars (or transducers) are able to capture complex entities and to consider the linguistic context of sequences. For instance, a duration expression is often preceded by a preposition such as *over*, *during* or *for*. The Unitex grammars allow us to specify this requirement more easily than with regular expressions[2].

We designed two graphs for the detection of negation markers that occur either on the left (e.g. *His work-up* ***ruled out*** *an acute coronary event*) or on the right (e.g. *Genetic testing* ***has not yet been performed***) of a medical entity and five graphs identifying each TIMEX3 attribute. Each transducer produces a lexically relevant output that is integrated as features into our CRF model.

[2]The star in Figure 1 indicates the end of the left context, meaning that this part of the grammar is used for computing matches but is not extracted (i.e. a preposition such as *over* must precede the article *the* but will not be included in the TIMEX3+DURATION entity).

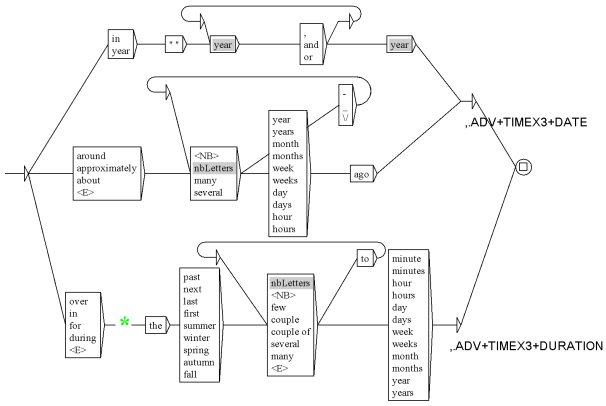

Figure 1: Example of Unitex transducer for the extraction of TIMEX3+DATE and TIMEX3+DURATION expressions

Events	P	R	F1
Max	0.915	0.891	0.903
ES:$\langle span \rangle$	0.892	0.878	0.885
Median	0.887	0.846	0.874
Max	0.900	0.875	0.887
EA:Polarity	0.870	0.857	0.864
Median	0.868	0.813	0.839

Table 3: Results for the EVENT subtasks

4 Results

In this section, we present our results on the final test data. We trained two different models: one was trained on the "Train" subcorpus, the other one on the "Train" and "Dev" datasets. We expected better results with the second model, given the larger training corpus. As the performance gain was quite limited between the two models, we do not mention here the performances of the first model.

In order to provide some context to these results, we also report the median and maximum results. Table 3 shows results for the EVENT tasks (ES, EA:polarity) and results for the TIMEX3 tasks (TS, TA) are reported in Table 4.

There are many more EVENT entities than TIMEX3 entities in the whole corpus (Train, Dev and Test). Consequently, we assume that if a class is more represented in the training corpus, the per-

Timex3	P	R	F1
Max	0.840	0.758	0.795
TS:⟨*span*⟩	0.777	0.564	0.653
Median	0.779	0.539	0.637
Max	0.815	0.735	0.772
TA:Class	0.752	0.545	0.632
Median	0.755	0.499	0.618

Table 4: Results for TIMEX3 subtasks

TA:attributes	P	R	F1
TA:date	0.768	0.573	0.656
TA:time	0.481	0.144	0.222
TA:prepostexp	0.979	0.814	0.889
TA:quantifier	0.594	0.288	0.388
TA:set	0.742	0.441	0.554

Table 5: Detailed results for TIMEX3 attributes

formances obtained on this class tend to be better: as we observed, the performances obtained for the EVENT class are higher than the ones for the TIMEX3 class.

In Table 5, we detail our results for each of the TIMEX3 attributes. Our system obtained very good results for the PREPOSTEXP attribute, probably because the lexical variation in this class is quite limited. By contrast, the results for the TIME and QUANTIFIER attributes remain low: this is partially explained by their low representation in the training corpus – the lowest among all the TIMEX3 attributes.

5 Discussion

A careful error analysis of 20% of the test corpus enabled us to pinpoint strategies for further improvement. We only considered TIMEX3 entities for this analysis. We identified four main types of issues that we can easily fix. We report the proportions of the various types of errors in Table 6.

Positional errors First, we noticed that some "easy" temporal expressions such as *Spring 2010* or *September 14, 2010* are not tagged by our system, even if they are detected by our linguistic rules and are thus inserted into the CRF's matrix of features. In a significant number of cases, the expressions are at the beginning of the sentence. The tuning of the tagger according to a specific type of documents (newspapers) may partially explain these errors, and weighting the CRF's features would be a simple way of addressing this issue. Indeed, theses weights could favour some features instead of others. For instance, if we apply a greater weight onto the feature reporting the linguistic information given by the grammars, this feature will have a greater influence than the feature "position in the sentence" and the errors explained in this paragraph will be reduced.

Error type	Percent
Positional errors	14.6%
Improvable resources	13.5%
Graph inconsistency	5.3%
Corpora inconsistencies	49.1%
Other	17.5%

Table 6: Error analysis (TIMEX3)

Improvable linguistic resources Secondly, our system missed some entities because they are neither in the training corpus nor in our linguistic resources, for e.g. *in childhood* (TIMEX3+DATE), *QID* (TIMEX3+SET) and *short number* (TIMEX3+QUANTIFIER). One issue highlighted here is that our lexical coverage is still too narrow, and we thus need to further expand our linguistic resources. On the one hand, the Unitex grammars for TIMEX3 entities need to be extended in order to detect more – and more complex – expressions. For this purpose, a more detailed linguistic study of a medical corpus in English is needed. On the other hand, it would be beneficial to preprocess our terminological resources so that they would better reflect authentic medical texts. Some processes may consist in cleaning up UMLS entries with phrasings typical of classificatory systems (e.g. *not otherwise specified*), splitting long terms into smaller units and generation of lexical variants (e.g. *varices of the esophagus* → *esophageal varices*)[3].

Graph inconsistency Thirdly, some linguistic rules do not fit with the annotation scheme used in the reference corpora. Our DATE grammar, for example, is able to detect a date followed by a time (e.g. *September 2, 2010 at 18:45*) and associates the

[3]In this error class, we only quantified the errors due to the lack of completeness of our graphs, while the terminological resources are used for EVENT identification.

tag DATE to the whole sequence. However, it appeared that such sequences are actually split into a DATE entity and a TIME entity in the corpus. Fortunately, such minor problems can easily be corrected.

Corpora inconsistency We also noted repeated inconsistencies in the corpora that were provided. This is not surprising as the corpus was manually annotated and obtaining a satisfactory inter-annotator agreement can be complicated for such a task, due to the fastidious and time-consuming nature of manual annotation in general; for details, see the TimeBank Documentation (Pustejovsky et al., 2006). However, we believe that some inconsistencies tend to lower our results and we will discuss two of them.

The first one concerns incoherent annotation between the training and the test corpora and accounts for about 30% of the total error rate. For example, some sequences are tagged in the training corpus but not in the test corpus: consider the sequence *recently-diagnosed*, where the adverb *recently* was tagged as a DATE about 15 times in the training corpus, but is never annotated in the test corpus. This has an influence on our results as our system extracts all these kinds of sequences.

This class also includes errors rather due to a difference in perspectives. It seems that the human annotators have tried to capture relevant but maximally succinct entities. Indeed, few annotated sequences contain more than 2 words for the EVENT class, 3 for the TIMEX3 class. On the contrary, our rules are designed to recognize the longest sequences possible. As a consequence, they always capture modifiers like *approximately* or *at least* when occurring before a SET entity, i.e. *approximately 4 times a day* or *at least 3 months ago*. Such a delimitation is actually considered as incorrect during the evaluation procedure.

The second type of inconsistencies concerns the annotation of prepositions and determiners and also affects all the corpora; these errors represent about 20% of the total error rate. During the development phase, we noted that the boundaries of two identical entities are sometimes different because the preposition and/or the determiner is not always considered. For example, we observed three different annotations in the data for the sequence "at the same time", namely, *at the same time*, *the same time* and *same time*. We made the same observation for other prepositions introducing temporal expressions, i.e. *for*, *over* (DURATION) and *in*, which were inconsistently annotated. For consistency, we adjusted our system to include the determiner in the entity, but not the preposition. Since the evaluation only considered as correct entities with the same offsets as those of the manually annotated entities, such a variation among the reference annotations inevitably had a negative impact on our score.

Nonetheless, we think that establishing a common definition of the temporal adverbial before the annotation phase is needed. Indeed, the preposition is structurally and semantically important for the adverbial (Gross, 1990): *in September*, for example, has a DATE attribute, but *since September* is a DURATION expression. Moreover, adopting this adverbial configuration would address issues at higher levels of linguistic analysis. Thus, such a definition step seems to be really important.

Finally, the *Other* category contains occasional errors. For instance, the sequence *in less than 6-weeks* is tagged as a DATE in the reference corpus, but as a DURATION by our system: this tag is incorrect given the semantic context of the expression. We consider these cases as learning errors due to a lack of occurrences of such constructions. One way to minimize this problem would be to increase the number of texts used for training.

6 Conclusion

This paper presents a hybrid approach for medical and temporal information extraction. Our objective was to evaluate the scalability of tools that have been previously developed to process French news articles. Our methodology relies on a single CRF model enriched by linguistic knowledge, and has been slightly adapted. We reused the terminological component of the UMLS and designed high-precision linguistic rules that feed the statistical model. The results are encouraging given the time constraints, but could definitely be improved using different strategies, some of which are detailed regarding major classes of errors.

References

S. Bethard, G. Savova, W.-T. Chen, L. Derczynski, J. Pustejovsky, and M. 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.

M. Constant and A. Sigogne. 2011. Mwu-aware part-of-speech tagging with a crf model and lexical resources. In *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World*, MWE '11, pages 49–56, Stroudsburg, PA, USA. Association for Computational Linguistics.

D. De Meyere, T. Klein, T. François, B. Fauquert, J.-C. Debongnie, C. Radulescu, N. Mbengo, M. Ouro Koura, Y. Coppieters 't Wallant, and C. Fairon. 2015. Automatic annotation of medical reports using SNOMED-CT: a flexible approach based on medical knowledge databases. In *Proceedings of the 7th Language Technology Conference (LTC2015)*, pages 519–523.

M. Gross. 1990. *Syntaxe de l'adverbe*, volume 3 of *Grammaire transformationnelle du français*. ASSTRIL.

J. D. Lafferty, A. McCallum, and F. C. N. Pereira. 2001. Conditional random fields : probabilistic models for segmenting and labeling sequence data. pages 282–289.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Sébastien Paumier. 2003. *De la reconnaissance des formes linguistiques à l'analyse syntaxique*. Ph.D. thesis, Marne-la-Vallée.

J. Pustejovsky, J. Littman, R. Saurí, and M. Verhagen. 2006. Timebank 1.2 documentation.

J. Pustejovsky, K. Lee, H. Blunt, and L. Romary. 2010. Iso-timeml: An international standard for semantic annotation. *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*.

W. Styler, S. Bethard, S. Finan, M. Palmer, S. Pradhan, P. de Groen, B. Erickson, T. Miller, C. Lin, G. Savova, and J. Pustejovsky. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

W. Sun, A. Rumshisky, and O. Uzuner. 2013. Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics*, 46, Supplement:S5 – S12.

P. Watrin, L. de Viron, D. Lebailly, M. Constant, and S. Weiser. 2014. Named entity recognition for german using conditional random fields and linguistic resources. *Workshop Proceedings of the 12th KONVENS 2014*.