Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era
J. Delgado et al. (Eds.)

doi:10.3233/SHTI210811

Extraction of Temporal Structures for Clinical Events in Unlabeled Free-Text Electronic Health Records in Russian

Anastasia A. FUNKNER[a,1] Dmitrii A. ZHURMAN[a] and Sergey V. KOVALCHUK[a,b]
[a] *ITMO University, Saint Petersburg, Russia*
[b] *National Almazov Medical Research Centre, Saint Petersburg, Russia*

Abstract. The important information about a patient is often stored in a free-form text to describe the events in the patient's medical history. In this work, we propose and evaluate a hybrid approach based on rules and syntactical analysis to normalise temporal expressions and assess uncertainty depending on the remoteness of the event. A dataset of 500 sentences was manually labelled to measure the accuracy. On this dataset, the accuracy of extracting temporal expressions is 95,5%, and the accuracy of normalization is 94%. The event extraction accuracy is 74.80%. The essential advantage of this work is the implementation of the considered approach for the non-English language where NLP tools are limited.

Keywords. time expression extraction; normalization; syntactical parsing; corpus; clinical text mining; machine learning

Extraction of temporal expressions from electronic health records (EHR) helps restore the chronology of the patient's diseases and order all his/her events on a timeline. Extracted temporal expressions and their events make data more findable, interoperable, and reusable according to FAIR principles [1]. There have been four competitions for the extraction of temporary structures in clinical texts [2]. However, most methods and approaches are not applicable for clinical texts in Russian because of the lack of labelled corpora [3]. Previously we developed an unsupervised approach to extract sentences with explicit temporal expressions but that approach has its drawbacks: imprecise retrievable constructions and difficulty in assessing obtained results [4].

Firstly, it is necessary to implement the extraction of temporal expressions (TEs) from sentences using rule-based methods. These TEs should be normalized to a single format (YYYY-MM-DD). Secondly, sentences with TEs should be parsed the syntactical parsers. Thirdly, we need to find a path from the defined TE to the right event in the syntactic tree. The algorithm to extract events is shown in Figure 1a. We compare common syntactic parsers for Russian and choose DeepPavlov because of its regular updates and detailed documentation. In this implementation, we use Spacy for writing rules because it shows a higher processing speed (35 sentences per sec upon 7 sentences per sec in Yargy). We develop 260 rules for TEs detection in Russian. For normalization, we used ready-made Python libraries dateparser and rutimeparser.

[1] Corresponding Author, ITMO University, Kronverksky Pr. 49, bldg. A, St. Petersburg, 197101, Russia; E-mail: funkner.anastasia@itmo.ru.

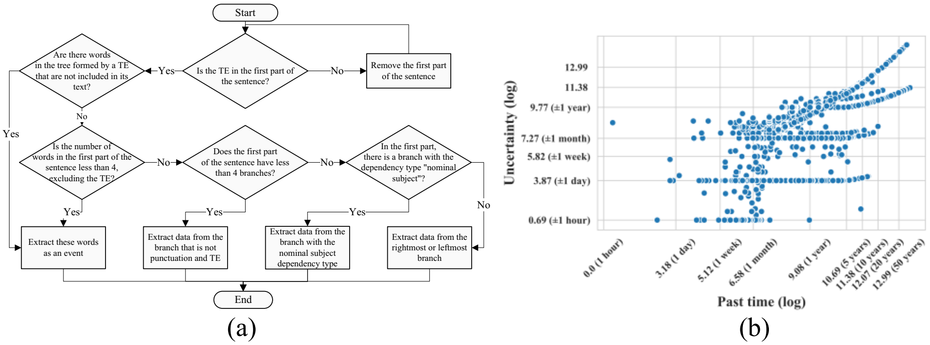
Start
Is the TE in the first part of the sentence?
Yes
No
Remove the first part of the sentence
Are there words in the tree formed by a TE that are not included in its text?
Yes
No
Is the number of words in the first part of the sentence less than 4, excluding the TE?
No
Does the first part of the sentence have less than 4 branches?
No
In the first part, there is a branch with the dependency type "nominal subject"?
Yes
Yes
Yes
No
Extract these words as an event
Extract data from the branch that is not punctuation and TE
Extract data from the branch with the nominal subject dependency type
Extract data from the rightmost or leftmost branch
End
(a)
Uncertainty (log)
12.99
11.38
9.77 (±1 year)
7.27 (±1 month)
5.82 (±1 week)
3.87 (±1 day)
0.69 (±1 hour)
0.0 (1 hour)
3.18 (1 day)
5.12 (1 week)
6.58 (1 month)
9.08 (1 year)
10.69 (5 years)
11.38 (10 years)
12.07 (20 years)
12.99 (50 years)
Past time (log)
(b)

Figure 1. Methods and results: (a) the algorithm for events extraction from a syntactic tree, and (b) logarithmic uncertainty for retrieved events.

The Cardiology Research Institute (Tomsk, Russia) provides anonymized EHRs, consisting of events and patient information. The Research Institute dataset includes 7777 sentences with temporal expressions (7277 and 500 for train and test sets). On the manually labelled test dataset, the accuracy of extracting TEs is 95.5%, the accuracy of normalization is 94%, and the events extraction accuracy is 74.8%.

We apply a trapezoidal membership function with a remoteness parameter to assess the uncertainty of extracted events. Figure 1(b) shows the uncertainty (log-log scale) for 6344 events. As can be seen, events of one category (known day, only month or year) form distinct line patterns. These lines bend as the age ratio increases linearly. Event uncertainty shows how reliable can be extracted events. Uncertainty scores can be used to build more accurate models as an additional feature.

In this paper, we propose an approach for extracting temporal structures and events in the absence of labelled data corpora for medical texts. With proper syntactic parsers, the models can be implemented for any other language. As the approach is focused on working without a labelled corpus, we believe it could find broader application in other languages with a lack of available public corpora and NLP tools in the medical domain.

Acknowledgements.

This research is financially supported by The Russian Science Foundation, Agreement #19-11-00326.

References

[1] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:1–9.

[2] Tang B, Wu Y, Jiang M, Chen Y, Denny JC, Xu H. A hybrid system for temporal information extraction from clinical text. J Am Med Informatics Assoc 2013;20:828–35.

[3] Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical Natural Language Processing in languages other than English: Opportunities and challenges. J Biomed Semantics 2018;9:1–13.

[4] Funkner AA, Kovalchuk S V. Time Expressions Identification Without Human-Labeled Corpus for Clinical Text Mining in Russian. Comput. Sci. -- ICCS 2020, Springer International Publishing; 2020, p. 591–602.