Creating an Annotated Health Record Dataset in a Limited-Resource Environment

Kristof Anetta

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, Brno, Czech Republic
xanetta@fi.muni.cz

Abstract. This paper demonstrates a workflow for creating a dataset of annotated electronic health records in an environment that is limited in terms of both language resources and expert availability. From preannotation using rule-based methods to the redundancy of multiple annotators per document and the resulting degrees of confidence for each annotation, including the possible avenues of data augmentation in order to be able to train large language models, this paper discusses the practical considerations of how to make the best of the resource-strapped situation shared by so many researchers who analyze health records.

Keywords: Electronic health records, EHR, annotation, named entity recognition, NER, medical concept mining.

1 Introduction

The lack of annotated data is a notorious issue in the field of electronic health record (EHR) analysis. The free-text data of electronic health records is widely considered to be a valuable yet largely untapped resource containing information about both medical science and the populations involved. However, the data exists in a form that cannot be properly understood by common large language models (LLMs) due to their being trained on natural language, not the dense, domain-specific, abbreviated structure of health record text.

While there are powerful LLMs for biomedical text in the English language (such as Gatortron [11] by NVIDIA and the University of Florida, and many others [10]), the situation in small languages such as Czech is dire and not likely to improve in the near future. This is due to the fact that there are no publicly available databases of health records and very few have been made available even to research teams. Adding to that, Czech does not have a large representation in the Unified Medical Language System (UMLS), making even vocabulary-based methods difficult. From the point of view of medical language processing, Czech can be considered a low-resourced language, and just like in so many other languages of similar size, there is no easy way of computationally locating medical concepts in free text - it needs to be annotated using human

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2023, pp. 157–164, 2023.

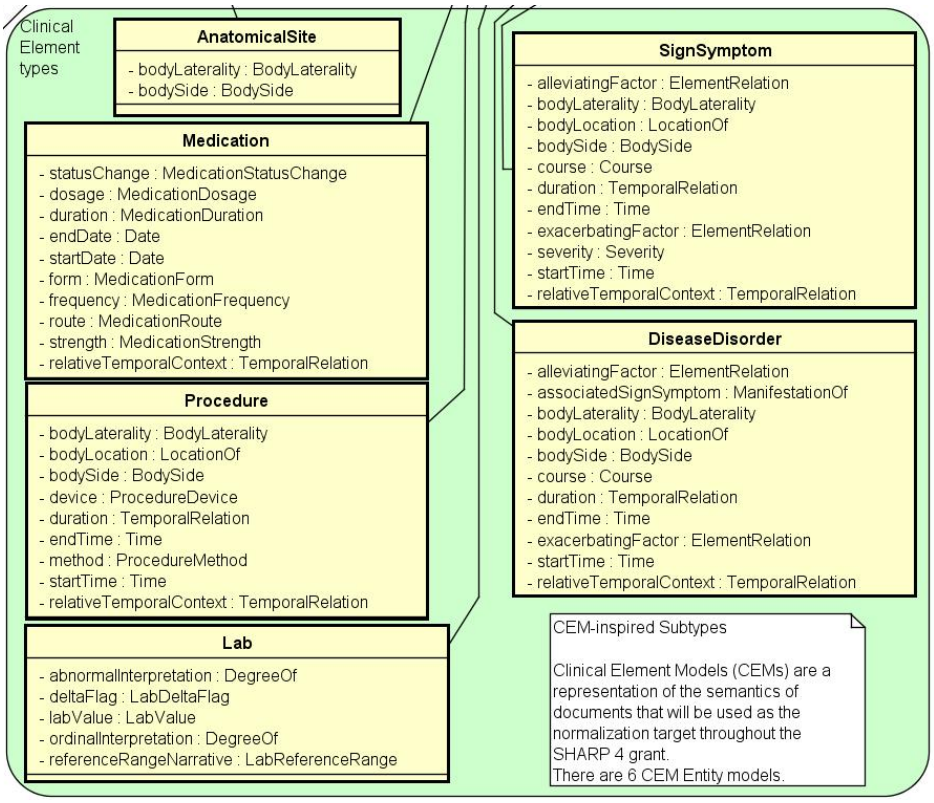

Fig. 1: Core entities in the type system of Apache cTAKES [1].

labor, which is slow and expensive, especially considering the amounts of data needed to train large Transformer-based models.

This paper presents a workflow that can, despite adverse conditions in terms of resources, lead to the creation of annotated health record data of reasonable quality.

2 Preparatory considerations

2.1 Selection of health records for annotation

Corpora of health records often contain distinct categories of medical text ranging from fluent narrative to almost tabular representation of laboratory values. To have enough training material for the dominant text types, but at the same time to be able to cover most of them, a reasonable sampling approach would reflect the ratios present in the whole corpus.

In this project, data for annotation was selected from a corpus of Czech health records collected at the Masaryk Memorial Cancer Institute in Brno, Czech Republic, totaling more than 42 million words in over 150,000 records detailing the stories of more than 4,200 patients. A balanced subset of 168 records, just under 50,000 words, was selected for human annotation.

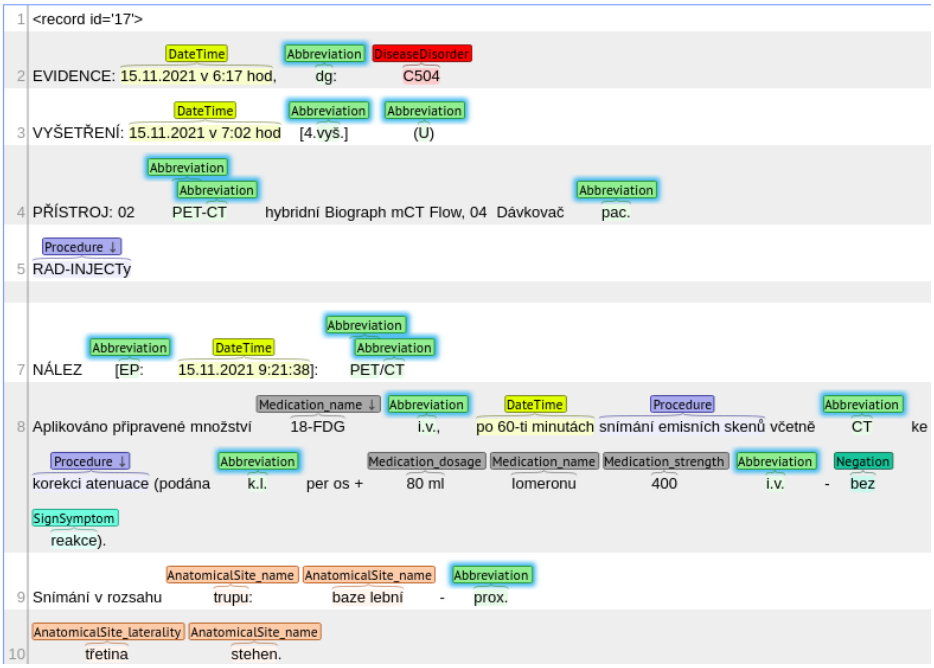

Fig. 2: BRAT tool interface.

2.2 Deidentification of records

Depending on the confidentiality clearance of recruited annotators, deidentification of texts can either be carried out before they are handed to the annotators, or it can be a part of the annotators' task.

In this project, all records selected for annotation were both automatically and manually searched for the occurrence of person names and other identifiers - the students tasked with annotation received safely deidentified data.

2.3 Choice of annotation schema

In order for the results to be commensurable with other research, the chosen annotation schema should be based on a standard already used in the field. This project is based on the six core clinical elements [1] (Figure 1) in the type system of Apache cTAKES [5], a major open-source NLP system for the extraction of clinical information from free text: *AnatomicalSite, DiseaseDisorder, Lab, Medication, Procedure, SignSymptom*.

To represent a few additional practical categories relevant for the medical domain, *Abbreviation, DateTime, and Negation* were added (Apache cTAKES represents these in a different way), and several core clinical elements were expanded into multiple annotation types to reflect some of the elements' deeper attributes: *AnatomicalSite_name, AnatomicalSite_laterality, Lab_name, Lab_unit, Lab_value, Medication_dosage, Medication_name, Medication_strength*.

Fig. 3: BRAT annotation dialog with the option of entering confidence and abbreviation expansion.

3 Workflow

3.1 Technology

In this project, the BRAT annotation tool [6] was chosen for its lightweight versatility, transparent and reusable file formats, and the option of recording comments and degrees of confidence for each annotation. Figure 2 shows the BRAT annotation interface and Figure 3 shows the annotation dialog box with the options available to annotators.

3.2 Preannotation

To maximize the efficiency of human annotators, any category of medical concepts that can be reliably annotated using rule-based methods should be preannotated before human annotators receive the text. However, these annotations should be editable so that after they are verified by the human, they become authoritative annotations, of a status equal to that of the manually entered ones.

In this project, preannotation vocabularies were compiled for

- names of medications registered in the Czech Republic, using the public database of the State Institute for Drug Control [7]
- common medical abbreviations, merging several available lists [4,2,3]

For an even more thorough preannotation, it is advisable to design regular expressions capturing repetitive character patterns such as

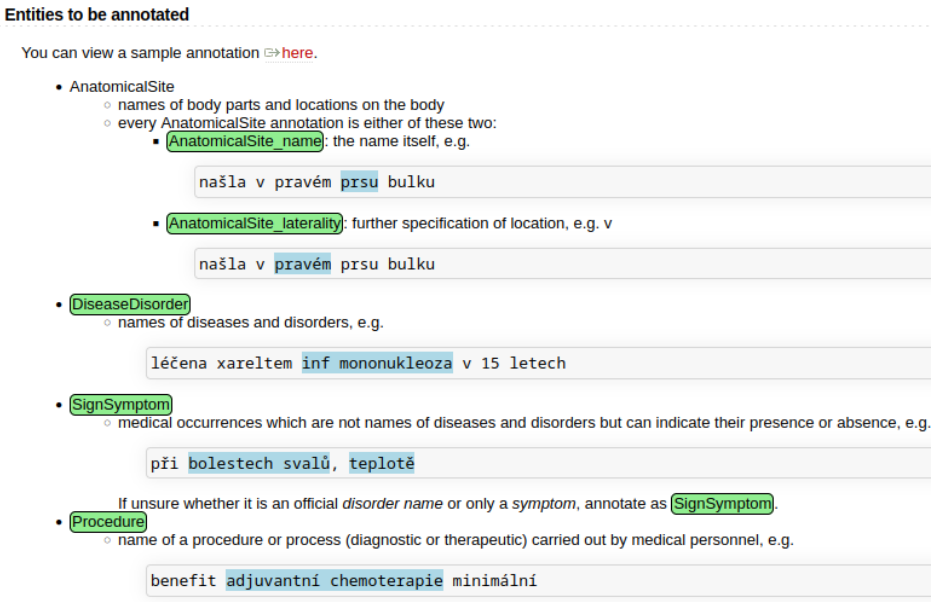

Fig. 4: Examples from the annotators' manual.

– quantities and units ("15mm")
– time expressions ("15:22")
– drug dosage regimen ("1-0-1")

and others, if these fit into the chosen annotation schema.

3.3 Human annotation process design

The situation of limited resources often includes the unavailability of experts whose annotation can be considered gold standard without reservation. While it would be enough to have one expert annotate each record, in the more common scenarios where the annotating workforce is only partially qualified or its qualification consisted in a short training, there is reasonable motivation to have multiple annotators per record.

The rationale for this is that multiple-person annotation produces both a high confidence "consensus set" of annotations, where the simultaneous decision of multiple humans to annotate a particular string raises its confidence almost to gold standard level, and also a wide and varied "fuzzy set" of annotations only entered by one of the annotators, which may or may not be perfectly correct, but are still highly valuable for training large language models' entity recognition (after all, automatic medical NER performing as well as a less qualified human annotator would be a grand achievement).

In this project, 11 university students were recruited. Each student received instruction in the form of an annotators' manual (see Figure 4), which explained the technical process of annotating in BRAT and introduced the types of annotations the students were expected to enter. Students were also instructed to revise preannotations and correct them if necessary. Since individual strings can

fall into multiple medical categories, multiple different annotations of the same string were allowed.

5 sub-datasets of just under 10,000 words were created and 2 or 3 annotators were assigned to annotate each of these sub-datasets. To mitigate issues such as failure to complete the task or serial position effect, records were shuffled for each annotator so the starting and ending positions were different for everyone.

Table 1 shows the numbers of annotations acquired in this project while following this workflow.

4 Prospects of LLM training

50,000 words is not enough training material to fine-tune a large Transformer model. But there are multiple options of how it can help create a sufficient amount of data.

4.1 Iterative augmentation

The limited-resource environment dataset of annotations can be used as the basis for a data augmentation process by a series of bootstrapping cycles with the following structure:

1. Training a NER model using the annotated data available in step *n*
2. Using this model to annotate a larger unannotated dataset planned for step *n+1*
3. Human-reviewing representative amounts of the resulting annotation and tweaking the *n+1* annotations, e.g. programmatically removing repeated patterns of incorrect annotation
4. Producing a final set of annotations for the *n+1* dataset
5. Repeating this process with an even larger dataset, using dataset *n+1* as *n*

This approach is similar to the work of [9], visualized in Figure 5.

Table 1: Annotation count at different stages of the annotation workflow.

Stage	**Annotation count**
Initial state of health records	0
Rule-based preannotations	4,266
Preannotations handed to annotators	9,368
New annotations entered by annotators	22,798
Total number of human-verified or human-entered annotations	32,166
Total number of tokens with human-verified or human-entered annotation	45,032

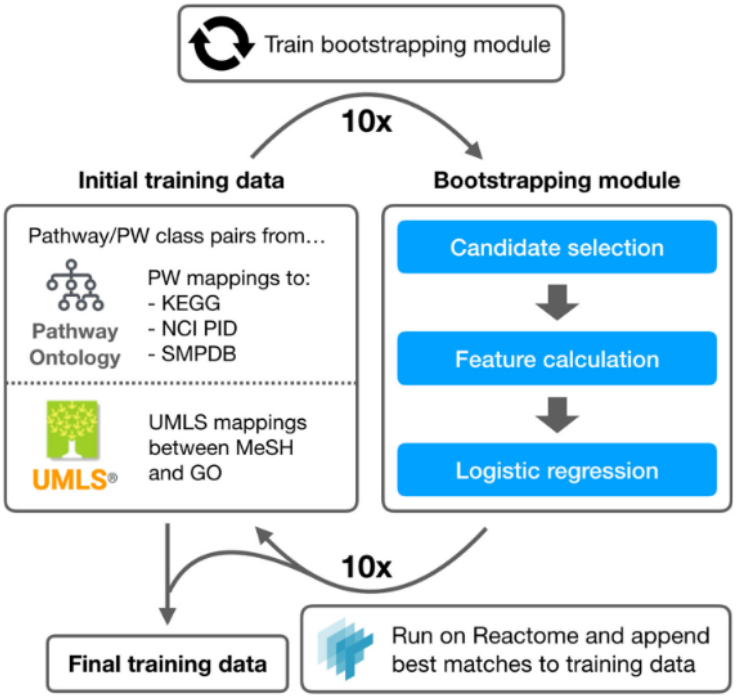

Fig. 5: Visualization of the bootstrapping procedure of [9], similar to the one proposed here.

4.2 Data synthesis

Another way of increasing the volume for training large models is to augment the data using a combination of the following:

- Synonym replacement - creating copies of annotated data while replacing human-annotated concepts with different concepts of the same type, e.g. using an external vocabulary of signs and symptoms (UMLS or other) to create variations of original sentences that contain a *SignSymptom* annotation.
- LLM-assisted synthesis of data similar to the original data, varying sentence structures and interchanging entities. This kind of approach recently gained popularity thanks to the fast growth of publicly available large language models. A notable example of a synthetic health record generation approach using prompts for LLMs can be found in [8].

5 Conclusion

It is apparent that this approach introduces many imperfections into the data along the way. However, in limited-resource scenarios, imperfect data is still infinitely better than none. As long as systems trained using such data are used in the capacity of assisting doctors with decisions and making their data more readable, they might easily have a net positive effect, but they first need to be created and evaluated in terms of what they are good for.

This paper serves to demonstrate one of the possible approaches where limited human and data resources are gradually developed into a usable dataset, and to encourage researchers in a similar situation to get inspired by it.

Acknowledgements. The work in this paper was carried out within the project MUNI/G/1763/2020: *AIcope – AI Support for Clinical Oncology and Patient Empowerment*. The analyzed Czech data was kindly provided by the Masaryk Memorial Cancer Institute in Brno, Czech Republic.

References

1. Apache cTAKES - User FAQs — svn.apache.org. `https://svn.apache.org/repos/infra/websites/production/ctakes/content/user-faqs.html`, [Accessed 09-11-2023]
2. Institute of Endocrinology: List of abbreviations – Institute of Endocrinology — endo.cz (2009), [Accessed 19-10-2023]
3. Karviná-Ráj hospital: List of abbreviations – Karviná-Ráj hospital — nspka.cz (2017), [Accessed 19-10-2023]
4. Jiráková, P.: List of the most common medical abbreviations – Alfabet — alfabet.cz (2014), [Accessed 19-10-2023]
5. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: 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 (2010)
6. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: BRAT: a web-based tool for NLP-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 102–107 (2012)
7. SÚKL, State Institute for Drug Control: Medicinal Products Database (in Czech Databáze léčivých přípravků DLP) — opendata.sukl.cz (2023), [Accessed 19-10-2023]
8. Tang, R., Han, X., Jiang, X., Hu, X.: Does synthetic data generation of llms help clinical text mining? arXiv preprint arXiv:2303.04360 (2023)
9. Wang, L.L., Thomas Hayman, G., Smith, J.R., Tutaj, M., Shimoyama, M.E., Gennari, J.H.: Predicting instances of pathway ontology classes for pathway integration. Journal of biomedical semantics **10**, 1–11 (2019)
10. Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., Pfeffer, M.A., Fries, J., Shah, N.H.: The shaky foundations of large language models and foundation models for electronic health records. npj Digital Medicine **6**(1), 135 (2023)
11. Yang, X., Chen, A., PourNejatian, N., Shin, H.C., Smith, K.E., Parisien, C., Compas, C., Martin, C., Flores, M.G., Zhang, Y., et al.: Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. arXiv preprint arXiv:2203.03540 (2022)