Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230281

Information Extraction from Medical Texts with BERT Using Human-in-the-Loop Labeling

Hendrik ŠUVALOV[a,1] and Sven LAUR[a] and Raivo KOLDE[a]
[a] *University of Tartu, Estonia*
ORCiD ID: Hendrik Šuvalov https://orcid.org/0000-0001-9625-3552,
Sven Laur https://orcid.org/0000-0002-9891-3347,
Raivo Kolde https://orcid.org/0000-0001-9625-3552

Abstract. Neural network language models, such as BERT, can be used for information extraction from medical texts with unstructured free text. These models can be pre-trained on a large corpus to learn the language and characteristics of the relevant domain and then fine-tuned with labeled data for a specific task. We propose a pipeline using human-in-the-loop labeling to create annotated data for Estonian healthcare information extraction. This method is particularly useful for low-resource languages and is more accessible to those in the medical field than rule-based methods like regular expressions.

Keywords. BERT, information extraction, natural language processing, medical texts, named entity recognition

1. Introduction

Low-resource languages, such as Estonian, do not have much available annotated data which makes training neural network language models difficult [1]. We use human-in-the-loop labeling to annotate data for finding specific entities like drug names, disease names or classifications etc. [2]. First, we use a naive regex with as high recall as possible to get samples with more frequent positive examples, then we start annotating the data, train the model and use its predictions to annotate the rest of the data faster. We used this method to fine-tune a model pre-trained on unlabeled Estonian medical data to extract cancer *TNM* stages[2] from free text and got remarkable results.

2. Methods

We used a naive regex to enrich the initial texts with positive examples of *TNM* stages. We then started annotating the data and after every 50 annotations, we used a BERT model [3] pre-trained on Estonian electronic health records and fine-tuned it with all the labeled data. We then used its predictions to find the positive examples to annotate faster.

[1] Corresponding Author: Hendrik Šuvalov, E-mail: hendrik.suvalov@ut.ee.
[2] *TNM* stages - https://www.cancer.gov/about-cancer/diagnosis-staging/staging

With each new iteration of training, we saw how accurately it predicted the stages and stopped when we were convinced it had learned the task.

3. Results

We evaluated the performance of the model, taking our advanced regular expression that we use in our workflows as the ground truth. Our test set consisted of 10'000 examples with 1:10 positive example ratio. After the model was trained on 150 examples, among which 14 were positive examples, it started predicting *TNM* stages with precision of 0.914 and recall of 0.867 and after 500 examples, it achieved precision of 0.820 and recall of 0.951. We manually analyzed the false positives annotated by the model of which there were 200 and found that 38 of them were actually correct, meaning the BERT based tagger caught the cases, but our regular expressions did not. The code is available at https://github.com/HealthInformaticsUT/labelstudio-ner-tagger.

4. Discussion

Extracting *TNM* stages is a relatively simple task that does not require many annotations for the model to learn and is not very context-dependent, which is one of BERT's strengths in comparison to rule-based approaches [3]. For more ambiguous cases, such as extracting mentions of family history, it could be much more efficient. Also, selecting which instances to train with or annotate can be approached in many ways, for example clustering similar texts or correcting the annotations for cases where the model is not very confident [4].

5. Conclusions

Pre-trainable neural network language models are useful for information extraction tasks but require annotated data, which low-resource languages often lack [1]. Human-in-the-loop labeling is an effective method for generating these annotations and the resulting models trained on this data can produce impressive results, often extracting cases that commonly used rule-based methods miss [2].

References

[1] Liu Z, Jiang F, Hu Y, Shi C, Fung P. NER-BERT: A pre-trained model for low-resource entity tagging. arXiv [csCL] [Internet]. 2021 [cited 2023 Jan 9], doi: https://doi.org/10.48550/arXiv.2112.00405

[2] Zhao Y, Liu J. Human-in-the-loop based named entity recognition. In: 2021 International Conference on Big Data Engineering and Education (BDEE). IEEE; 2021. p. 170–6, doi: https://doi.org/10.1109/BDEE52938.2021.00037

[3] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional Transformers for language understanding. arXiv [csCL] [Internet]. 2018 [cited 2023 Jan 10], doi: https://doi.org/10.48550/arXiv.1810.04805

[4] Zhang Y, Nie A, Zehnder A, Page RL, Zou J. VetTag: improving automated veterinary diagnosis coding via large-scale language modeling. NPJ Digit Med [Internet]. 2019;2(1):35