Proceedings of the TERM21 Workshop , pages 30–32
Language Resources and Evaluation Conference (LREC 2022), Marseille, 20-25 June 2022

Evaluating Pre-Trained Language Models for Focused Terminology Extraction from Swedish Medical Records

Oskar Jerdhaf[1], Marina Santini[2], Peter Lundberg[1], Tomas Bjerner[1], Yosef Al-Abasse[1], Arne Jönsson[3], Thomas Vakili[4]
[1]Linköping University Hospital, [2]RISE Research Institutes of Sweden,
[3]Linköping University, [4]Stockholm University

Abstract

In the experiments briefly presented in this abstract, we compare the performance of a generalist Swedish pre-trained language model with a domain-specific Swedish pre-trained model on the downstream task of *focused terminology extraction* of implant terms, which are terms that indicate the presence of implants in the body of patients. The fine-tuning is identical for both models. For the search strategy we rely on KD-Tree that we feed with two different lists of term seeds, one with noise and one without noise. Results shows that the use of a domain-specific pre-trained language model has a positive impact on focused terminology extraction only when using term seeds without noise.

Keywords: terminology extraction, implant terms, generalist BERT, domain-specific BERT

1. Introduction

Intuitively, a domain-specific pre-trained language model should preform better than a generalist pre-trained language model when the downstream task is domain specific. However, this commonsense intuition is not always confirmed by empirical results (Gu et al., 2021; von der Mosel et al., 2021; Zheng et al., 2022). Since the effect of domain-specific pre-trained language models on domain-specific downstream tasks is not fully investigated, in the experiments presented here we further explore this issue. Building domain-specific pre-trained language models is expensive and the implication is that each different domain should then have its own specific pre-trained language model. Obviously, if generalist pre-trained language models perform competitively, considering using generalist models rather than domain-specific models would become a strong option in order to save time and money. We delve more into this issue and we explore the downstream task of focused terminology extraction. Focused terminology extraction indicates the extraction of a relatively small family of specific terms, i.e. terms that represent a specialized semantic field. In this case we focus on the extraction of terms that indicate or suggest the presence of “implants” in electronic medical records (EMRs) written in Swedish.

2. Evaluating Terminology Models

The evaluation of Automatic Terminology Extraction (ATE) models is notoriously difficult. As pointed out during the latest shared task competition at TermEval2020: ”Taking into account the unpredictability of many machine learning approaches and the considerable variety between the potential outputs, as demonstrated in this shared task, it is essential for ATE to be evaluated beyond precision, recall, and f1-scores” (Rigouts Terryn et al., 2020). Evaluation is even more difficult in the absence of domains or sub-domains where gold standards are not available. This situation is very common when dealing with the specialized terms that characterize focused terminology extraction. In this case, the terms candidates must be evaluated by domain experts **on the output of focused terminology extraction systems**. With this type of evaluation, that we call *posterior evaluation*, we will only know the number of good candidate terms (yes-terms), bad candidate terms (no-terms) and terms where the annotators feel ”unsure”, but we remain unaware of the total numbers of good, bad and unsure terms in the whole corpus.

In previous experiments (Jerdhaf et al., 2021)[1], we built a initial gold standard based on the posterior evaluation of a generalist Swedish pre-trained language model. We say ”initial” because the gold standard will be incrementally augmented in the way we explain in Section 4. The gold standard for this task has been designed with three categories, namely **yes-terms** (good candidates), **no-terms** (bad candidates) and **u-terms** (unsure and ambiguous terms). This gold standard is the manually evaluated output of a focused terminology extraction model that was fine-tuned on the generalist Swedish KB-BERT model (Malmsten et al., 2020) to discover implant terms unsupervisely. Top ranked candidate implant terms were presented to domain experts (two MRI physicists) for manual evaluation. Results were promising according to our experts. However, we observed that the number of candidate terms that were NOT indicative of implants was quite high. Therefore, we decided to investigate whether a pre-trained domain-specific language model would help in decreasing the number of bad candidates.

[1] The research has been approved by the Swedish Ethical Review Authority (Etikprövningsmyndigheten), authorization number: 2021-00890 to Peter Lundberg.

	Gold Standard	KB-BERT		SweDeClin-BERT	
Term seeds	-	Term seeds w/ noise	Term seeds w/o noise	Term seeds w/ noise	Term seeds w/o noise
YES-terms	1267	648	409	383	575
NO-terms	2930	1503	796	723	73
Discoveries	-	2868	4018	2807	1279
Total	4197	5019	5223	4036	1927

Table 1: Breakdown of terms extracted by the models and the overlap with terms in the gold standard.

3. Data and Datasets

The data used for the downstream task are medical records written in Swedish. We use the medical records of two clinics (cardiology and neurology) that belongs to the LIU-Hospital-EMRs-collection, described in Jerdhaf et al. (2021).

4. Method

The aim of the experiments described below is to compare a focused terminology extraction model fine-tuned on the generalist Swedish pre-trained KB-BERT (Malmsten et al., 2020) with a focused terminology extraction model fine-tuned on the domain-specific (clinical) Swedish pre-trained SweDeClin-BERT (Vakili et al., 2022) on the extraction of implant terms.

Both models have been fine-tuned using the same parameters on the same dataset created from the medical records of two clinics (cardiology and neurology). For the search strategy, we used KDTree (Python, sklearn.neighbors.KDTree) (Pedregosa et al., 2011) with two different lists of term seeds, one with noise (753 terms) and one without noise (1267 implant terms) (see example in Figure 1, right hand-side). Term seeds play a very important role in this type of modelling because they are used to generate random queries. This means that for each term seed, a sentence containing the term was randomly chosen from the dataset and used to find contextually similar sentences. The similarity of contextually similar sentences is based on word embeddings. Essentially, the model will select candidate terms that have a similar role and position as the term seeds of the queries. At this stage of our research the creation of the queries is randomized. This randomization has the advantage of discovering new candidates (that we call *discoveries*) at each run of the model. Discoveries are the terms brought to surface by the randomized queries. The role of discoveries is paramount since it is unthinkable and unfeasible that two or more MRI physicists read millions of medical records and annotate implant terms in one go. In our approach, at each run, the domain experts will be presented new discoveries that, when annotated, will increase the gold standard. An example of how the domain-experts annotate the discoveries is shown in Figure 1, left hand-side. It is a iterative process that will repeat until the majority of discoveries will be in the Yes-term list of the gold standard. It is important to notice that the models will always surface new discoveries because medical records will be added to the current collection over time and because new implant artefacts will be placed on the market and used on patients. What we want to achieve at this point of our research is to identify the model that: 1) maximize the number of good candidate terms already present in the Yes-term list of the gold standard; 2) minimize the number of bad candidate terms already present in the No-term list of the gold standard; 3) return a number of discoveries that when evaluated have the same distribution pattern as described in points 1 and 2, i.e. many good candidates and few bad candidates.

5. Results and Evaluation

According to the results shown in Table 1, the focused terminology extraction model fine-tuned on the domain-specific (clinical) Swedish pre-trained SweDeClin-BERT in combination with term seeds without noise (Column 6) meets the expectations stated in points 1 and 2 of the previous section .

In order to verify the 3rd expectation, we handed over the 1279 discoveries generated by that model to two domain experts. Manual evaluation of the 1279 discoveries meets our expectation as formulated in point 3 because the two domain experts agreed on assessing 750 Yes-terms and they also agreed on rating 91 No-terms. They had discordant ratings on the rest. We observe that the distribution trend of the Yes- and No-terms of the manually evaluated discoveries matches the trend of the Yes- and No-terms found in the gold standard.

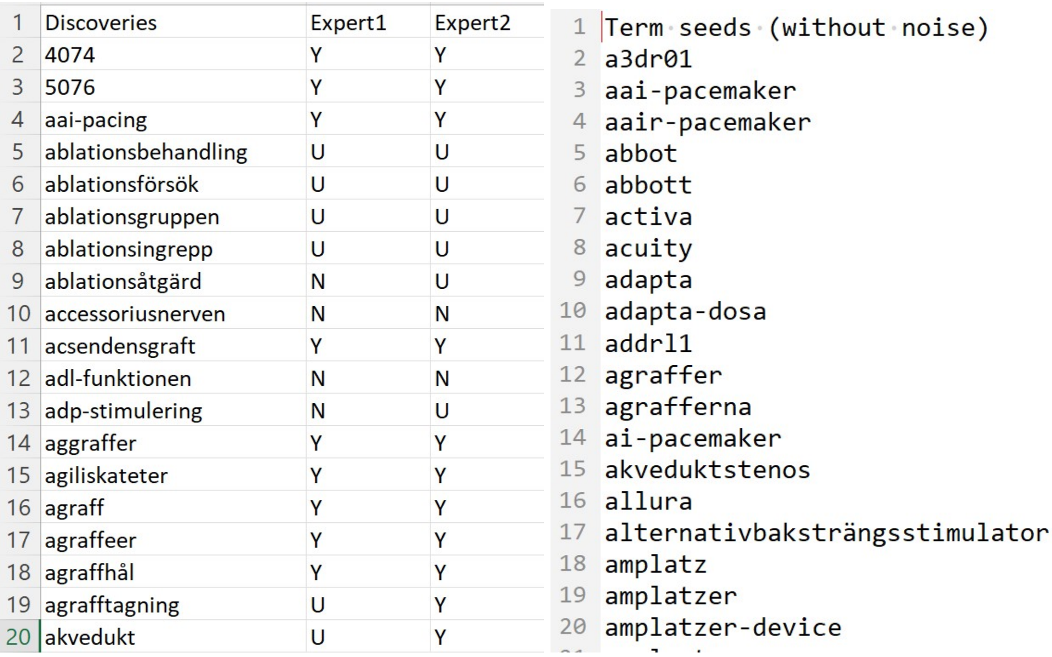
6. Discussion

Results shows that the use of a domain-specific pre-trained language model has a positive impact on focused terminology extraction only when using term seeds without noise. This means that a domain-specific pre-trained model has a positive effect under certain conditions.

We are aware that the randomization of the queries as motivated in Section 4 has the downside of conflicting with the principle of experimental replicability. We are currently studying alternative solutions that allow diversification of the results and assure replicability.

7. Conclusion

In this abstract we shortly presented ongoing research on unsupervised focused terminology extraction. Although this is a difficult research area especially for the lack of well-established gold standards and evaluation metrics, results are encouraging. The current gold standard for this task is available for inspection and reuse.



Discoveries	Expert1	Expert2
4074	Y	Y
5076	Y	Y
aai-pacing	Y	Y
ablationsbehandling	U	U
ablationsförsök	U	U
ablationsgruppen	U	U
ablationsingrepp	U	U
ablationsåtgärd	N	U
accessoriusnerven	N	N
acsendensgraft	Y	Y
adl-funktionen	N	N
adp-stimulering	N	U
aggraffer	Y	Y
agiliskateter	Y	Y
agraff	Y	Y
agraffeer	Y	Y
agraffhål	Y	Y
agrafftagning	U	Y
akvedukt	U	Y

```
Term seeds (without noise)
a3dr01
aai-pacemaker
aair-pacemaker
abbot
abbott
activa
acuity
adapta
adapta-dosa
addrl1
agraffer
agrafferna
ai-pacemaker
akveduktstenos
allura
alternativbaksträngsstimulator
amplatz
amplatzer
amplatzer-device
```

Figure 1: Discoveries (left), term seeds without noise (right)

Acknowledgements

This research was funded by **Vinnova** (Sweden's innovation agency), **Grant number**: 2021-0169 and by the **DataLEASH** project.

References

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Jerdhaf, O., Santini, M., Lundberg, P., Karlsson, A., and Jönsson, A. (2021). Implant term extraction from swedish medical records–phase 1: Lessons learned. In *Swedish Language Technology Conference and NLP4CALL*, pages 35–49.

Malmsten, M., Börjeson, L., and Haffenden, C. (2020). Playing with words at the national library of Sweden–making a Swedish BERT. *arXiv preprint arXiv:2007.01658*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Rigouts Terryn, A., Hoste, V., Drouin, P., and Lefever, E. (2020). Termeval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (acter) dataset. In *6th International Workshop on Computational Terminology (COMPUTERM 2020)*, pages 85–94. European Language Resources Association (ELRA).

Vakili, T., Lamproudis, A., Henriksson, A., and Dalianis, H. (2022). Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data. (Accepted to LREC 2022), June.

von der Mosel, J., Trautsch, A., and Herbold, S. (2021). On the validity of pre-trained transformers for natural language processing in the software engineering domain. *arXiv preprint arXiv:2109.04738*.

Zheng, Z., Lu, X.-Z., Chen, K.-Y., Zhou, Y.-C., and Lin, J.-R. (2022). Pretrained domain-specific language model for general information retrieval tasks in the aec domain. *arXiv preprint arXiv:2203.04729*.