Implant Terms: Focused Terminology Extraction with Swedish BERT Preliminary Results

Oskar Jerdhaf[1], **Marina Santini**[2], **Peter Lundberg**[3,4], **Anette Karlsson**[3,4], **Arne Jönsson**[1]
[1] Department of Computer and Information Science, Linköping University, Sweden
oskje724@student.liu.se|arne.jonsson@liu.se
[2] RISE, Digital Health, Sweden
marina.santini@ri.se
[3] Center for Medical Image Science and Visualization (CMIV), Linköping University, Linköping, Sweden
[4] Department of Medical Radiation Physics and Department of Health, Medicine and Caring Sciences
Linköping University, Linköping, Sweden
Peter.Lundberg@liu.se|Anette.k.karlsson@regionostergotland.se

Abstract

Certain implants are imperative to detect before MRI scans. However, implant terms, like 'pacemaker' or 'stent', are sparse and difficult to identify in noisy and hastily written electronic medical records (EMRs). In this paper, we explore how to discover implant terms in Swedish EMRs with an unsupervised approach. To this purpose, we use BERT, a state-of-the-art deep learning algorithm, and fine-tune a model built on pre-trained Swedish BERT. We observe that BERT discovers a solid proportion of indicative implant terms.

1 Introduction

Domain-specific terminology extraction is an important task in a number of areas, such as knowledge base construction (Lustberg et al., 2018), ontology induction (Sazonau et al., 2015) or taxonomy creation (Šmite et al., 2014). We present an exploratory experiment on an underinvestigated type of terminology extraction that we call "focused terminology extraction". With this expression, we refer to terms or to a nomenclature that represent a specific semantic field. More specifically, we explore focused terminology related to the semantic field of terms that indicate or suggest the presence of "implants" in electronic medical records (EMRs) written in Swedish.

Implant terms are domain-specific words indicating artificial artefacts that replace, partially or in full, organs, bones, arteries or other parts of the human body. Common implants are devices such as 'pacemaker', 'shunt', 'prosthesis' or 'stent'.

It is important to know if a patient has an implant because MRI-scanning is incompatible with some implants (e.g. the 'pulmonary artery catheter') or maybe partially compatible with some of them (e.g. the 'mitraclip'). Unsafe implants must be considered before MRI-scanning, as they may be contraindicative, while conditional implants can be left in the patient's body, if conditions are appropriately accounted for. One of the safety measures in MRI-clinics is to ask patients whether they have or have had an implant. This routine is not completely reliable, because a patient (especially if elderly) might have forgotten about the presence of implants in the body. Arguably, referring physicians are aware of the constraints of specific implants and, prior to an MRI-examination, they should go through the patient's medical history by reading the patient's EMRs. EMRs are digital documents, but the information they contain is not structured or organized in a way that makes it trivial to find implant terms quickly and efficiently. This downside can be addressed by automatically trying to identify implant terms from the EMRs based on their contextual usage, e.g. using word embeddings. To this purpose, we use BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), which is the state-of-the art in computational linguistics and deep learning for many NLP tasks, such as text classification, question answering, sentiment analysis, Named Entity Recognition and Text Summarization. However, to our knowledge, BERT has never been used to detect distributional word similarity for a terminology extraction task. In the experiment described in this paper, we explore how Swedish BERT performs on this task and present preliminary results. The aim of the model is to find as many valid instances of implant-related words as possible in free-text (unstructured) EMRs. Results are encouraging and manual domain expert-based evaluation shows that BERT discovers a solid proportion of indicative implant terms.

2 Related Work

"Focused terminology" refers to the mentions of a relatively small number of technical terms. From a semantic perspective, focused terminology extraction is particularly challenging because the task implies an unsupervised discovery of a handful of specialized terms scattered in millions of words across unstructured textual documents, such as EMRs. EMRs are written by physicians who typically use a wide range of medical sublanguages that are not only based on regular medical jargon, but also include unpredictable word-shortening and abbreviations, spelling variants of the same word (including typos!), numbers, and the like. What is more, these sublanguages vary across hospitals and clinics.

Focused terminology extraction is still underexplored. Little work exists on this task, although its usefulness in real-world applications is extensive. In the experiment presented here, we build on research carried out at Linköping University in close cooperation with Linköping University Hospital. Kindberg (2019) started this exploration and relied on Word2Vec (Mikolov et al., 2013). In his experiments, carried out on the EMRs of the cardiology clinic (660 000 EMRs), out of the 500 terms, 340 (68%) were considered relevant. For the same task, Nilsson et al. (2020) used Swedish BERT (Malmsten et al., 2020) on 9,4 million sentences from the same cardiology clinic. The results presented in Nilsson et al. (2020) show that out of the 148 evaluated terms, 68 (46%) in their given context were assessed to be clearly indicative of implants or other harmful objects. 27 words (18%) were assessed to be possibly indicative for implants or other harmful objects in the contexts they appeared in, and 53 words (36%) were considered non-indicative.

It must be emphasized that the results by Kindberg (2019) and by Nilsson et al. (2020) are not directly comparable between them and with the results presented in this paper because different evaluation strategies and different evaluation metrics were used. However, although the experiment presented in this paper is based on a larger EMR corpus that includes two clinics, we capitalize on the knowledge created by these two previous experiences, in the way discribed in the next section.

3 Electronic Medical Records (EMRs)

The data used in our experiment is the text of EMRs from two clinics at Linköping University Hospital, namely the cardiology clinic and the neurosurgery clinic. These EMRs span over the latest five years and amount to about 1 million EMRs, when taken individually, and about 48 000 when grouped by unique patient. The size of the current corpus is bout 71 million words (see Table 1). Each EMR varies greatly in length, from just a few words to hundreds of words. The EMRs have not been fully anonymised, therefore we cannot release this data at the time of this publication. However, we will distribute secondary linguistic data, such as automatically created implant glossaries and lists of BERT terms on the project website.

Clinics	Words	SingleEMRs	GroupedEMRs
Cardio	45 780 055	664 821	34 044
Neuro	25 440 484	314 669	14 526
Total	71 220 539	979 490	48 088

Table 1: Number of words and EMRs per clinic

4 Method

Generally speaking, a BERT model uses a transfer learning approach, and it is pre-trained on a large amount of data. After learning deep bidirectional representations from unlabelled text, BERT can be further fine-tuned for several downstream tasks. In our experiment, we focus on the fine-tuning of focused terminology extraction.

4.1 Pre-Trained Swedish BERT

For our focused terminology extraction task, we used the pre-trained Swedish BERT model released by The National Library of Sweden (Malmsten et al., 2020)[1]. To provide a representative BERT model for the Swedish language, the model was trained on approximately 15-20 gigabyte of text (200M sentences, 3000M tokens) from a range of genres and text types including books, news and internet forums. The model was trained with the same hyperparameters as first published by Google and corresponded to the size of Google's base version of BERT with 12 so-called transformer blocks (number of encoder layers), 768 hidden units, 12 attention heads and 110 million parameters.

4.2 Fine-Tuning Swedish BERT for Focused Terminology Extraction

To fine-tune the pre-trained model, we relied on PyTorch (Paszke et al., 2019) using the Huggingface transformers library (Wolf et al., 2019) freely

[1]https://github.com/Kungbib/swedish-bert-models

available and ready to use. The EMRs and the pretrained model were fed into a Python script. The fine-tuning was done at Linköping University Hospital. For this experiment, the BERT model was fine-tuned using only the CPU on Intel Xeon to gauge the processing time with ordinary computing resources normally available to most users.

Since no previous studies are available for word similarity based on BERT, we relied on settings successfully used for other tasks and documented in the literature as our starting point. We fine-tuned for 3 epochs with a learning rate of 5e-5 and a batch size of 32, as in SQuAD v1.1 (Devlin et al., 2019). The block size was set to 64, which means that sequences with lesser than 64 tokens are padded to meet this length, and sequences with more than 64 tokens are truncated. A block size of 64 is based on observation of our current data. However, since we are awaiting for additional data, this value may change in future. The model was trained with MLM (Masked LM), a technique which allows bidirectional training. MLM consists in replacing 15% of the words in each sequence with a [MASK] token before feeding word sequences into BERT. The model then attempts to predict the original value of the masked words, based on the context provided by the other, non-masked, words in the sequence. The weights for the softmax classification layer were randomly initialized. All the other hyperparameters not mentioned here were set to the default. The fine-tuning with the settings described above took approximately 15 hours per clinic to complete.

4.3 Discovering Implant Terms

We used the MRI-safety handbook (SMRlink) publicly available at the hospital website to automatically create a glossary of implant or implant-related terms. The terms in the glossary were not verified by domain experts. Rather, on purpose, we left some noise (i.e. non-implant terms) in the glossary to assess the resistance to noise and the robustness of the whole approach with minimum preprocessing and no post-processing. In this experiment, we use a version of the glossary containing 753 terms (unigrams) including noise. With these terms, a keyword matching was conducted against the EMRs. From the results of the keyword matching, two separate data subsets were created. One of them containing only sentences without glossary terms and the other one containing sentences with at least one glossary term. We used the same representation scheme for both subsets. Then, for each glossary term, a set of 15 queries was created from the subset containing glossary terms. The queries were used to find terms in similar distributional contexts in the subset of sentences not containing glossary terms. We used KDTree (short for k-dimensional tree) to search the vector space for terms that appear in contexts that are similar to the queries. We decided that, given our data size, the pairwise cosine similarity metrics would have been too inefficient with ordinary computing resources. KDTree, on the other hand, is a binary space-partitioning data structure for organizing points in a k-dimensional space and it is useful when using multidimensional search key (e.g. range searches and nearest neighbour searches). In this experiment, we used the nearest neighbour search version of KDTree in Python (*sklearn.neighbors.KDTree*) (Pedregosa et al., 2011) with the default Minkowski distance. We extracted 7 nearest neighbours for a given term in the subset of sentences not containing glossary terms. We realized, however, that although DKTrees are normally efficient data structures, storing the vector values over millions sentences and their values in relation to each other inside the KDTree is extremely memory-intensive. In order to speed up this process, the data was split into chunks which were mapped into separate KDTrees. Each individual KDTree was used to generate results for all queries and then the most contextually similar words and sentences across all of the chunks were selected for the final results. The results used in this paper were generated with chunks of 50 000 tokenized sentences clustered in each KDTree.

4.4 Expert-Based Evaluation

A manual evaluation of BERT discoveries was carried out by two MRI-physicists, with different level of experience, from the Radiology clinic at Linköping University Hospital. We surmize that the difference in expertise or experience is important to pin down different degrees of familiarity with technical terms. For this reason, we decided to hand to the assessors only BERT terms without any context, i.e. without the full sentences from where the terms were extracted.

The two MRI-physicists received an excel file containing the list of terms to be assessed and short instructions (this excel file is available on

the project website). They were asked to assess the terms using the following ratings on a three-degree scale: **Y** = *yes, it gives me an indication that the patient has or has had an implant*; **N** = *No, it DOES NOT give me any indication that the patient has or has had an implant*; **U** = unsure, *the term could or could not give me an indication of an implant, but I cannot decide without more context*.

The evaluation was divided into two parts. The first part focused on the inter-rater agreement between the two domain experts, who assessed independently a subset of BERT discoveries, namely a sample of 813 (16,4%) out of 4951 BERT terms.

The second part focused on the assessment made only by one MRI-physicist, who evaluated a more extensive list of BERT terms. Results are presented in the next section.

5 Results and Discussion

We measured the inter-rater agreement between the two MRI-physicists by using percentage (i.e. the proportion of agreed items in relation to the whole without chance correction), the classic unweighted Cohen's kappa (Cohen, 1960) and Krippendorff's alpha (Krippendorff, 1980) to get a straightforward indication of the raters' tendencies.

Table 2: Inter-rater agreement on 813 BERT terms

Terms	Percentage	Cohen's Kappa	Krippendorff's Alpha
813	75.9%	0.6	0.597

Table 2 shows the breakdown of the inter-rater agreement. The raters agree on 617 terms, of which 248 are indicative implant terms. They tend to disagree most when they have to decide if term is NOT indicative or when they simply felt "unsure" (88 terms). Some terms were classified as "unsure" by one rater and indicative by the other rater (55 terms). Finally, on 53 terms the raters had completely divergent opinion: 36 terms were indicative for one rater, but not indicative for the other, while for 17 terms the assessment was reversed. Overall, the values in Table 2 show that both kappa and alpha coefficients are approx. 0.6, and both these values indicate a "moderate" agreement according to the magnitude scale for kappa (Sim and Wright, 2005), and the alpha range (Krippendorff, 2011). The moderate agreement between the two domain experts may suggest that selective experience and/or expertise could play a role in recognizing implant terms, and BERT terms can contribute in alerting professionals about the presence of implants that could otherwise be overlooked.

Table 3: Single rating of 4443 BERT terms

Terms	Y	N	U
4443	1470 (33%)	2603 (58.5%)	369 (8.3%)

The results of the rating by one MRI-physicist on 4443 terms (89,7%) out of 4951 is shown in Table 3. According to this rating, 33% of the BERT terms are indicative of an implant. This percentage was far beyond our expectation considering the size of the corpus and the noise both in the glossary and in the corpus. We think these results are encouraging since the BERT model presented in this paper is still exploratory and needs further refinements.

Undeniably the domain expertise is of fundamental importance for the refinement of the model, since the model sieve through extremely noisy textual data. The evaluation has helped us identify what kind of irrelevant words the model retrieves. Error analysis indicates that families of irrelevant words could be filtered out during pre-processing of EMRs. For instance: misspelled words in Swedish (e.g. *abltaion* and in English (e.g. *achive*), first name person noun (e.g. *Ann-Christin*) and general medical terms (e.g. *epidural*). The next step is then to filter out semantic families of words that create noise in the results.

6 Conclusion and Future Directions

In this paper, we presented preliminary results of a fine-tuned Swedish BERT model for focused terminology extraction. The model was devised to discover terms indicative of implant in Swedish EMRs. Although the task is challenging, manual evaluation of BERT terms presented without any context to the assessors reveals that the approach is rewarding, since a solid number of indicative terms were discovered by BERT regardless of noise in the glossary and the EMRs. These "discoveries" will be used to further refine the model in future experiments. In upcoming experiments, we will focus on a more systematic analysis of the hyperpameters' space when fine-tuning the model, as well as on the benefits (if any) of the Ball Tree search space to overcome the limitations of both cosine similarity and KDTrees.

Acknowledgements

This research was funded by Vinnova. Project title: Patient-Safe Magnetic Resonance Imaging Examination by AI-based Medical Screening. Grant number: 2020-00228.
Project Website: http://www.santini.se/mri-terms/

References

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Erik Kindberg. 2019. Word embeddings and patient records: The identification of mri risk patients. Bachelor's Thesis. Linköping University http://liu.diva-portal.org/smash/get/diva2:1324363/FULLTEXT01.pdf.

Klaus Krippendorff. 1980. Content analysis. *California: Sage Publications*, 7:l–84.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. Working Papers, University of Pennsylvania http://repository.upenn.edu/asc_papers/43.

Tim Lustberg, Johan Van Soest, Peter Fick, Rianne Fijten, Tim Hendriks, Sander Puts, and Andre Dekker. 2018. Radiation oncology terminology linker: A step towards a linked data knowledge base. *Studies in health technology and informatics*, 247:855–859.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of sweden–making a swedish bert. *arXiv preprint arXiv:2007.01658*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

A. Nilsson, J. Källbäcker, J. Monsen, L. Nilsson, M. Mattila, M. Jakobsson, and O. Jerdhaf. 2020. Identifying implants in patient journals using bert and glossary extraction. Student Report. Linköping University http://www.santini.se/mri-terms/2020-06-04_ProjectReportGroup1-729G81_Final.pdf.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8026–8037.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Viachaslau Sazonau, Uli Sattler, and Gavin Brown. 2015. General terminology induction in owl. In *International Semantic Web Conference*, pages 533–550. Springer.

Julius Sim and Chris C Wright. 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3):257.

Darja Šmite, Claes Wohlin, Zane Galviņa, and Rafael Prikladnicki. 2014. An empirically based terminology and taxonomy for global software engineering. *Empirical Software Engineering*, 19(1):105–153.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.