Focused Terminology Extraction for CPSs The Case of "Implant Terms" in Electronic Medical Records

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*—Language Technology is an essential component of many Cyber-Physical Systems (CPSs) because specialized linguistic knowledge is indispensable to prevent fatal errors. We present the case of automatic identification of *implant terms*. The need of an automatic identification of implant terms spurs from safety reasons because patients who have an implant may or may be not submitted to Magnetic Resonance Imaging (MRI). Normally, MRI scans are safe. However, in some cases an MRI scan may not be recommended. It is important to know if a patient has an implant, because MRI scanning is incompatible with some implants. At present, the process of ascertain whether a patient could be at risk is lengthy, manual, and based on the specialized knowledge of medical staff. We argue that this process can be sped up, streamlined and become safer by sieving through patients' medical records. In this paper, we explore how to discover implant terms in electronic medical records (EMRs) written in Swedish with an unsupervised approach. To this aim we use BERT, a state-of-the-art deep learning algorithm based on pre-trained word embeddings. We observe that BERT discovers a solid proportion of terms that are indicative of implants.**

I. Introduction

Domain-specific terminology extraction is an important task in a number of areas, such as knowledge base construction [1], ontology induction [2] or taxonomy creation [3].

We present experiments on an underexplored 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. The automatic identification and extraction of this kind of nomenclature are a common need in many domains, e.g. medicine, dentistry, chemistry, aeronautics, engineering and the like.

In these experiments, 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. More specifically, the aim of our experiments is to investigate whether it is possible to discover implant terms or implant-related words unsupervisely, i.e. without any labelled or annotated data. Results are evaluated by two domain experts. This task is part of an ongoing project at Linköping University Hospital, Sweden.

Implant terms are domain-specific words indicating artificial artefacts that replace or complement parts of the human body. Common implants are devices such as 'pacemaker', 'shunt', 'codman', 'prosthesis' or 'stent'. The need of an automatic identification of implant terms spurs from safety reasons because patients who have an implant may or may be not submitted to MRI scans. Magnetic resonance imaging (MRI) is very safe and most people are able to benefit from it. However, in some cases an MRI scan may not be recommended. Before having an MRI scan, the following conditions must be verified: (a) metal in the body and (b) being pregnant or breastfeeding. 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. When a patient has or is suspected to have an implant, the procedure of recognition and acknowledgement is manual, laborious and involves quite many human experts with specialized knowledge. The workflow of the current procedure is shown in Figure 1 and described in [4].

Even if implants have been removed, metallic or electronic parts (like small electrodes or metallic clips) may have been overlooked and left in situ, without causing harm to patient's health before the MRI. Normally, referring physicians may be aware of the limitation of specific implants, and prior to an MRI-examination, they should go through the patient's medical history by reading EMRs. EMRs are digital doc-

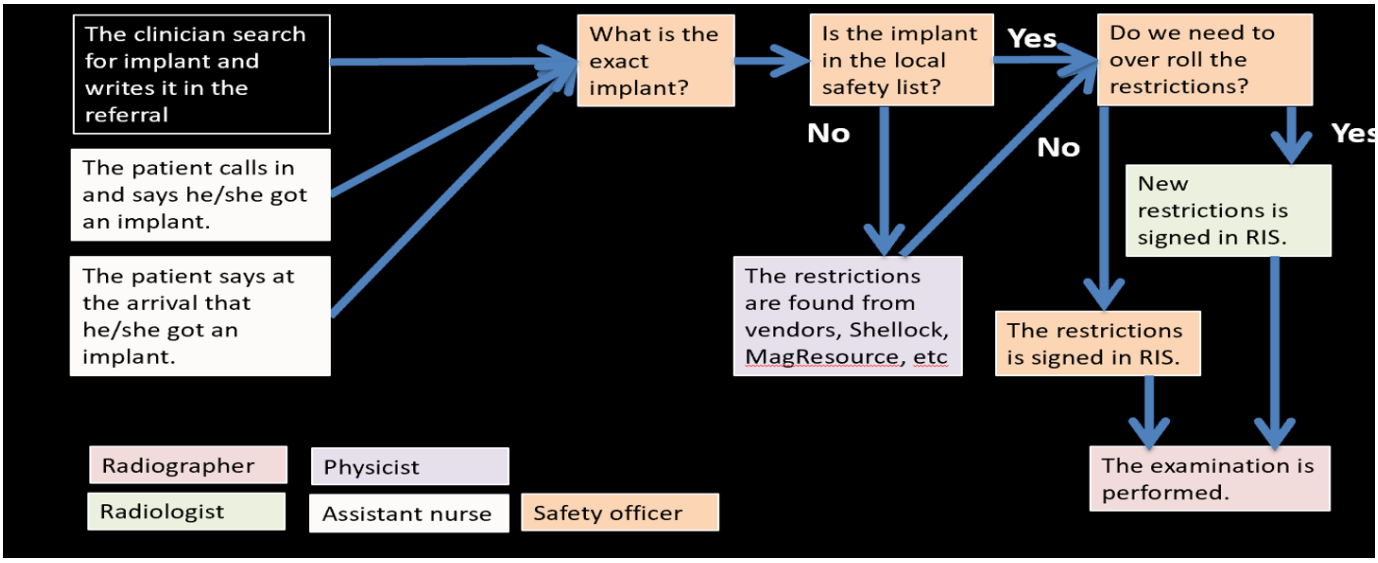

Fig. 1. Current workflow [4]

uments, 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 in EMRs based on their contextual usage, e.g. using word embeddings. In our experiments, we use BERT (Bidirectional Encoder Representations from Transformers) [5], which is the state-of-the art in computational linguistics and deep learning. The aim is to find as many validated instances of implant-related words as possible in free-text EMRs.

II. Related Work

Focused terminology extraction refers to mentions of a relatively small number of technical terms. From a semantic perspective, focused terminology extraction is challenging because the task implies the 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. Recent studies exist however on medical synonym discovery. For instance, [6] compare eight neural models on the task of finding disorder synonyms in English clinical free text. In their evaluation, ELMO models perform moderately better than the other models. In the experiments presented here we build on research carried out at Linköping University in close cooperation with Linköping University Hospital. [7] started this exploration and relied on Word2Vec. In his experiments, which were carried out on a small portion of EMRs, out of the 500 terms, 340 (68%) were considered relevant. For the same task, [8] used BERT, which currently guarantees unchallenged performances in many NLP tasks. The results presented in [8] show that out of the 148 evaluated terms, 68 (46%) were assessed to be clearly indicative of implants in their given context; 27 words (18%) were assessed to be possibly indicative of implants in the contexts they appeared in, and 53 words (36%) were considered non-indicative and not related to implants. It must be noticed that the results by [7] and by [8] are not directly comparable between them nor with our own experiments because different data samples and different evaluation methods were used. In our experiments, we build on these two previous experiences and apply BERT to a large EMR corpus that includes two clinics, neurosurgery and cardiology. We have already presented some preliminary results in [9].

III. Data: Electronic Medical Records

The data used in our experiments is the text of EMRs from two very different 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 48000 when groped by unique patient (the breakdown is shown in Table I). EMRs vary greatly in length, from just a few words to hundreds of words. From this data, two subcorpora have been created, one called “cardio” and the other called “neuro”. The two corpora used in these experiments have not yet been fully anonymised, therefore we are unable to release them at the time of this publication. However, we will distribute secondary linguistic data, such as automatically created wordlists on the project website.

IV. Method: BERT

Previous methods to represent features as vectors have been unable to capture the context of individual words in the

TABLE I
NUMBER OF WORDS AND EMRS PER CLINIC

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

texts, sometimes leading to a poor representation of natural language. When using a traditional text classifier, one of the simplest ways to represent text is to use bag-of-words (BOW), where each word (feature) in the text is stored together with their relative frequency, ignoring the position of the word in the sentence and in the text. A more advanced way to represent features is by using word embeddings, where each feature is mapped to a vector of numbers. The pioneer of this approach was a method called Word2Vec [10]. However, one limitation of this or of similar approaches was the mono-directionality of the system that could read only one side of the surrounding context. A big leap forward was achieved with BERT (Bidirectional Encoder Representations from Transformers), which uses a multi-headed self-attention mechanism to create deep bidirectional feature representations, able to model the whole context of all words in a sequence. Bidirectional refers to the ability of simultaneously learning left and right word context. Up to BERT, bidirectionality could be achieved only by modeling two separate networks for each direction that would later be combined, as in [11]. A BERT model uses a transfer learning approach, where 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 these experiments, we relied on PyTorch (an open source machine learning framework[1]) [12] and used the Huggingface transformers library for BERT [13] available and ready to use[2].

A. Swedish BERT

1) Pre-Trained Model: The pre-trained BERT model used in these experiments is the *bert-base-swedish-cased* released by The National Library of Sweden [14][3]. 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.

A BERT model has a predefined vocabulary. This vocabulary is a set of words known to the model and it is used to tokenize words. A token can in this case be a common word, a common subpart of a word or a single letter. Each object in the vocabulary of the model has a known embedding. To use the model for finding the embedding of a new word the model was used to tokenize the word, which means that it would try to rebuild the word using as few tokens from the vocabulary as possible. The pre-trained BERT-model used in this study had a vocabulary of 50325 words. Pre-trained model hyperparameters are listed in Table II.

TABLE II
PRE-TRAINING PARAMETERS

Hyperparemeter	Dimensions/Value
Dropout	0.1
Hidden Activation	GELU
Hidden Size	768
Embedding Size	512
Attentional Heads	12
Hidden Layers	12
Forward Size	3072
Vocabulary Size	50325
Trainable Parameters	$11 \cdot 10^7$

2) Fine-Tuning the Pre-Trained Model: The model was fine-tuned using the Adam algorithm with default values for its hyperparameters as indicated by [15]. The pre-processed EMRs and the pre-trained model were fed into a Python script.

TABLE III
PARAMETERS USED FOR FINE-TUNING

Hyperparameter	Dimension/Value
Epochs	3
Batch Size	32
Block Size	64
Learning Rate	$5e-5$

The model was trained for three epochs with MLM (Masked Language Model), a train batch size of 32, a 5e-5 learning rate and a block size of 64 (see Table III). These decisions were made partly based on the original BERT paper [5] and partly on [16]. The hyper-parameters not mentioned here were set to the default value. The fine-tuning took approximately 15 hours per clinic to complete using the computing resources shown in Table IV.

TABLE IV
DETAILS OF COMPUTING RESOURCES

Label	Description
CPU	Intel Xeon - 12x(E5-2620 v3)
GPU	NVIDIA Quadro M4000 [8GB(VRAM)|20GB(Shared)]
Clock Speed	2.40GHz
Memory (RAM)	40GB

3) Discovering Contextually-Similar 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. In these experiments, we use a version of the glossary containing 753 terms. With the 753

[1]https://pytorch.org/
[2]https://huggingface.co/transformers/
[3]https://github.com/Kungbib/swedish-bert-models

terms, queries were created. Here we present the results of a BERT model built using 15 queries for each term.

The queries were subsequently used to find words with similar contextual features in the corpus. This was done using the scikit-learn implementation of the KD-Tree algorithm [17]. Storing the vector values over a million sentences, and their values in relation to each other inside the KD-Tree is extremely memory-intensive. In order to speed up this process, the data was split into chunks which were mapped onto separate KD-Trees. Each individual KD-Tree 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 50000 tokenized sentences clustered together in each KD-Tree. A total of 4636 terms were discovered by BERT in this way.

B. Evaluation

To judge whether a term –without any context– is relevant and indicative of implants or other objects that could be harmful during an MRI-scan, special domain knowledge is required. In some cases, it may be obvious that a term indicates an implant. In other cases, it may be less obvious due to very specific-domain language, abbreviations, and other domain-related, or even clinic-related, ways of writing EMRs or describing certain phenomena. For this reason, a manual evaluation of BERT discoveries was carried out by two MRI-physicists from the Radiology clinic at Linköping University Hospital, who assessed independently the set of 4636 terms discovered by BERT.

The two MRI-physicists, received an excel file containing the list of terms to be assessed and short instructions. They were instructed to judge whether the term can give an indication that the patient has or has had an implant. They were asked to use 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 inter-rater agreement was then computed on their judgement to understand how indicative BERT terms are to professionals. Results are presented in the next section.

V. Results

We measured the inter-rater agreement between the two MRI-physicists by using percentage (i.e. the proportion of agreed upon documents in relation to the whole without chance correction), the classic unweighted Cohen's kappa [18] and Krippendorff's alpha [19] to get a straightforward indication of the raters' tendencies.

Cohen's kappa assumes independence of the two coders and is based on the assumption that "if coders were operating by chance alone, we would get a separate distribution for each coder" [20]. This assumption intuitively fits our expectations. Krippendorff's alpha is similar to Cohen's kappa, but it also takes into account the extent and the degree of disagreement between raters [20].

TABLE V
INTER-RATER AGREEMENT ON 4636 BERT TERMS

Terms	Percentage	Cohen's Kappa	Krippendorff's Alpha
4 636	75%	0.575	0.573

TABLE VI
BREAKDOWN BY RATER

Rater	Y	N	U
Rater-1	1 426 (30.8%)	2 701 (58.2%)	509 (11%)
Rater-2	1 321 (28.5%)	2 395 (51.5%)	920 (20%)

Concordant Assessment		
y	y	*1088*
u	u	*224*
n	n	*2163*
		3475
Discordant Assessment		
y	u	157
y	n	76
u	y	234
u	n	462
n	y	104
n	u	128
		1161

Fig. 2. Breakdown: concordant/discordant assessments by the two raters.

Tables V and VI show the breakdown of the inter-rater agreement of the 4636 terms discovered by BERT. Overall, the values in Table V shows that both kappa and alpha coefficients are approx. 0.57, and both these values indicate a "moderate" agreement according to the magnitude scale for kappa [21], and the alpha range [22]. The moderate agreement between the two domain experts may suggest that selective experience and/or expertise could play a role in recognizing implant terms (see Table VI), and BERT terms can contribute in alerting professionals about the presence of implants that could otherwise be overlooked. Essentially, BERT helped discover 23.5% of positive implant terms on which the two raters agree. This percentage was far beyond our expectation and we think it is encouraging since the BERT model presented in this paper is a pre-study. The raters agree on 3475 terms, of which 1088

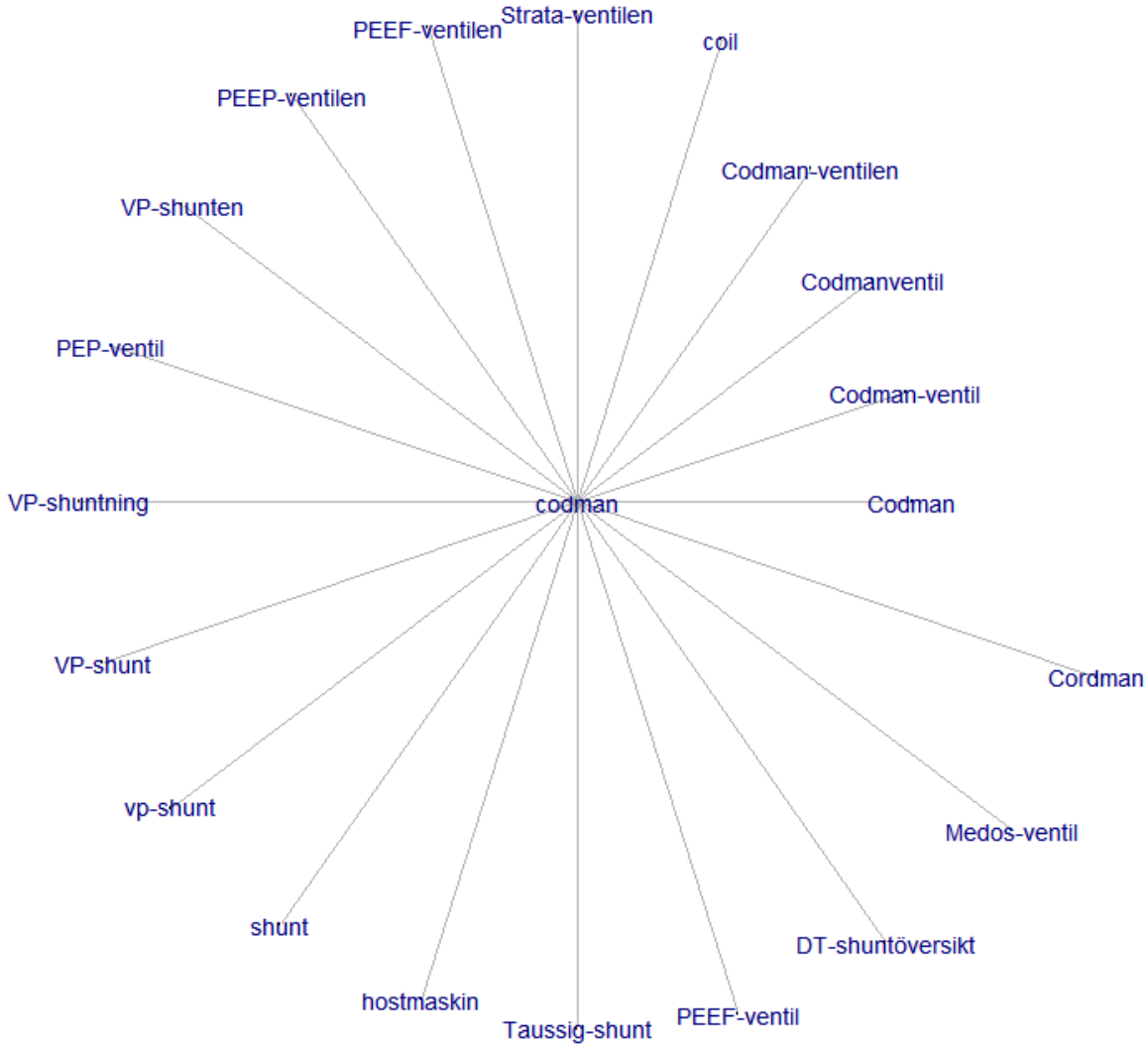

Fig. 3. BERT terms contextually similar to 'codman' (lowercase, center). The graph must be read anti-clockwise, starting from 'Codman' (capitalized). The length of the edges represents the distance of a term from the query term at the center. The figure shows that BERT can capture a wide variety of terms indicative of implants including the misspelled 'Cordman'.

are indicative implant terms (approx. 23.5%), 2163 terms are not indicative of implants, and for 224 terms the raters were "unsure". They disagreed on 1161 terms. This means that BERT helped discover 75% of terms on which the two raters are concordant (i.e. 1088+2163+224), and 25% on which they are discordant (see Figure 2). Out of the 1088 indicative BERT terms, about 900 were not in the glossary. Therefore these BERT terms make a useful addition to glossary expansion. Out of 2163 non-indicative BERT terms, about 2000 were not in the glossary, which means that the level of noise in the automatically created glossary is relatively small.

VI. Discussion

Undeniably the domain expertise is of fundamental importance for the refinement of the model, since the model sieve through extremely noisy textual data. The domain-expert evaluation has helped us identify what kind of irrelevant words the model retrieves. Error analysis indicates that families of irrelevant words were not filtered out during preprocessing. 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.

Although EMRs are noisy texts, BERT successfully discovers a good portion of useful implant terms. Figure 3 shows BERT terms that are contextually similar to the term 'codman' and give an intuitive idea of the relations between the terms. Codman is the product name of an implant[4]. Figure 3 must

[4]https://www.jnjmedicaldevices.com/en-US/codman-pumps/patient-support

be read anti-clockwise starting from capitalized 'Codman', which is the most contextually similar term to 'codman'. The most distant similar term is a mispelled version of the same word, i.e. 'Cordman'. Identifying misspellings or unpredictable abbreviations that abound in noisy texts like EMRs is certainly a benefit. In the same figure, we can see that BERT returns useful terms, such as the many variations of the word 'ventil' and 'shunt', but also terms that signal the presence of implant, but are not implant names, e.g. 'coil'. The raters agreed unanimously that all the terms in Figure 3 were terms indicative of implants with the exception of 'hostmaskin' (en: cough assist machine) that was assessed by both as non indicative of implants.

VII. Conclusion

In this paper we presented results of a BERT model for focused terminology extraction. The model is fully automated and was devised to discover terms indicative of implants in EMRs. Although the task is challenging, manual evaluation reveals that the approach is rewarding, since a solid number of indicative terms were discovered by BERT. These discoveries will be used to further refine the model in future. What is more, we have started the annotation of EMRs to create a supervised classification model that can identify patients who are MRI-incompatible, patients who are MRI-compatible and finally patients who are in the grey zone, i.e. cases that must be further investigated by MRI physicists. This annotation is manual and it is a painstakingly time-consuming work that can be carried out only by MRI experts. We are going to help our experts by highlighting the implant terms discovered by BERT in the EMRs themselves, so that the annotation can be sped up. The creation of annotated resources is a fundamental step to train or to evaluate advanced AI-based models.

Acknowledgement

This research was funded by **Vinnova** (Sweden's innovation agency), https://www.vinnova.se/

Project title: *Patient-Safe Magnetic Resonance Imaging Examination by AI-based Medical Screening*.

Call: Start your AI-journey! For public organizations.

Grant number: 2020-00228 to Peter Lundberg.

The study was approved by the Swedish Ethical Review Authority, Dnr 2021-00890 (to PL).

Project Website: http://www.santini.se/mri-terms/

References

[1] T. Lustberg, J. Van Soest, P. Fick, R. Fijten, T. Hendriks, S. Puts, and A. Dekker, "Radiation oncology terminology linker: A step towards a linked data knowledge base." *Studies in health technology and informatics*, vol. 247, pp. 855–859, 2018.

[2] V. Sazonau, U. Sattler, and G. Brown, "General terminology induction in owl," in *International Semantic Web Conference*. Springer, 2015, pp. 533–550.

[3] D. Šmite, C. Wohlin, Z. Galviņa, and R. Prikladnicki, "An empirically based terminology and taxonomy for global software engineering," *Empirical Software Engineering*, vol. 19, no. 1, pp. 105–153, 2014.

[4] J. Kihlberg and P. Lundberg, "Improved workflow with implants gave more satisfied staff," in *SMRT 28th Annual Meeting 10-13 May 2019*, 2019.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, 2019, pp. 4171–4186.

[6] E. Schumacher and M. Dredze, "Learning unsupervised contextual representations for medical synonym discovery," *JAMIA open*, vol. 2, no. 4, pp. 538–546, 2019.

[7] E. Kindberg, "Word embeddings and patient records: The identification of mri risk patients," B.Sc. thesis, Linköping University, 2019.

[8] A. Nilsson, J. Källbäcker, J. Monsen, L. Nilsson, M. Mattila, M. Jakobsson, and O. Jerdhaf, "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, 2020.

[9] O. Jerdhaf, M. Santini, P. Lundberg, A. Karlsson, and A. Jönsson, "Implant terms: Focused terminology extraction with swedish bert - preliminary results," in *Eighth Swedish Language Technology Conference (SLTC2020). Organised by University of Gothenburgå, Sweden*, 2020.

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[11] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 2227–2237.

[12] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8026–8037.

[13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, pp. arXiv–1910, 2019.

[14] M. Malmsten, L. Börjeson, and C. Haffenden, "Playing with words at the national library of sweden–making a swedish bert," *arXiv preprint arXiv:2007.01658*, 2020.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv*, pp. arXiv–1412, 2014.

[16] F. Li, Y. Jin, W. Liu, B. P. S. Rawat, P. Cai, and H. Yu, "Fine-tuning bidirectional encoder representations from transformers (bert)–based models on large-scale electronic health record notes: An empirical study," *JMIR medical informatics*, vol. 7, no. 3, p. e14830, 2019.

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

[18] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[19] K. Krippendorff, "Content analysis," *California: Sage Publications*, vol. 7, pp. l–84, 1980.

[20] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," *Computational Linguistics*, vol. 34, no. 4, pp. 555–596, 2008.

[21] J. Sim and C. C. Wright, "The kappa statistic in reliability studies: use, interpretation, and sample size requirements," *Physical therapy*, vol. 85, no. 3, p. 257, 2005.

[22] K. Krippendorff, "Computing krippendorff's alpha-reliability," http://repository.upenn.edu/asc_papers/43, 2011.