Implant Term Extraction from Swedish Medical Records – Phase 1: Lessons Learned

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

We present the case of automatic identification of "implant terms". Implant terms are specialized terms that are important for domain experts (e.g. radiologists), but they are difficult to retrieve automatically because their presence is sparse. The need of an automatic identification of implant terms spurs from safety reasons because patients who have an implant may be at risk if they undergo Magnetic Resonance Imaging (MRI). At present, the workflow to verify whether a patient could be at risk of MRI side-effects is manual and laborious. We claim that this workflow can be sped up, streamlined and become safer by automatically sieving through patients' medical records to ascertain if they have or have had an implant. To this aim we use BERT, a state-of-the-art deep learning algorithm based on pre-trained word embeddings and we create a model that outputs *term clusters*. We then assess the linguistic quality or term relatedness of individual term clusters using a simple intra-cluster metric that we call *cleanliness*. Results are promising.

1 Introduction

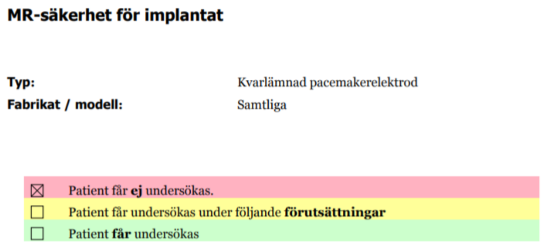
Domain-specific terminology extraction is an important task in a number of areas, such as knowledge base construction (Lustberg et al., 2018), ontology induction (Sazonau et al., 2015) or taxonomy creation (Šmite et al., 2014).

We present 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 specialized 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. learning from unlabelled data. This task is currently part of an ongoing project carried out together with LIU University Hospital. We present here the results and the lessons learned from Phase 1 of the project.

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 Magnetic Resonance Imaging (MRI). MRI scans are very safe and most people are able to benefit from it. However, in some cases an MRI scan may not be recommended. Before undergoing an MRI scan, the following conditions must be verified: (a) the presence of metal in the body and (b) being pregnant or breastfeeding. Implants are often metallic objects, therefore 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'). An example of a recommendation on implants is shown in Figure 1. The translated (narrative) version of the recommendation reads: "If a pacemaker electrode is present in the patient's body, then the patient cannot be exposed to MRI scanning".



MR-säkerhet för implantat

Typ:	Kvarlämnad pacemakerelektrod
Fabrikat / modell:	Samtliga

☒ Patient får **ej** undersökas.
☐ Patient får undersökas under följande **förutsättningar**
☐ Patient **får** undersökas

Figure 1: According to this recommendation, a patient having a pacemaker electrode in the body cannot undergo MRI scanning.

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 2 and described in (Kihlberg and Lundberg, 2019).

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 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 the terms from the EMR based on their contextual usage, e.g. using word embeddings. In our experiments, 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 several downstream tasks, e.g. text classification, question-answering or natural language understanding. Our downstream task is to find as many validated instances of implant-related words as possible in free-text EMRs. Here we present the lessons we have learned from Phase 1 of the project.

2 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 an unsupervised discovery of a handful of specialized terms scattered in millions of words across unstructured textual documents, such as EMRs. This characterization has some similarities with the "relevant but sparse" definition in Ittoo and Bouma (2013). 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 general medical synonym discovery. For instance, Schumacher and Dredze (2019) compare eight neural models on the task of finding disorder synonyms in English clinical free text. In their evaluation, ELMO models performs moderately better than the other models. Before the neural revolution and the word embeddings paradigm, models for synonym extraction have been proposed for many languages and also specifically for the Swedish language. The models for Swedish presented in Henriksson et al. (2012) are based on (by now) traditional word space models, namely Random Indexing and Random Permutation. The models were designed to identify both synonyms and abbreviations. These models were built on the Stockholm EPR Corpus (Dalianis et al., 2009) and synonym extraction was evaluated on the Swedish version of MeSH and its extension[1]. Results were encouraging, but limited to terms included in Swedish MeSH, which does not cover the whole medical terminology and, what is more, does not include graphical variations that are present in the informal medical sublanguage often used in medical records.

Focussed terminology extraction could be interpreted as a special case of Named Entity Recog-

[1] https://mesh.kib.ki.se/

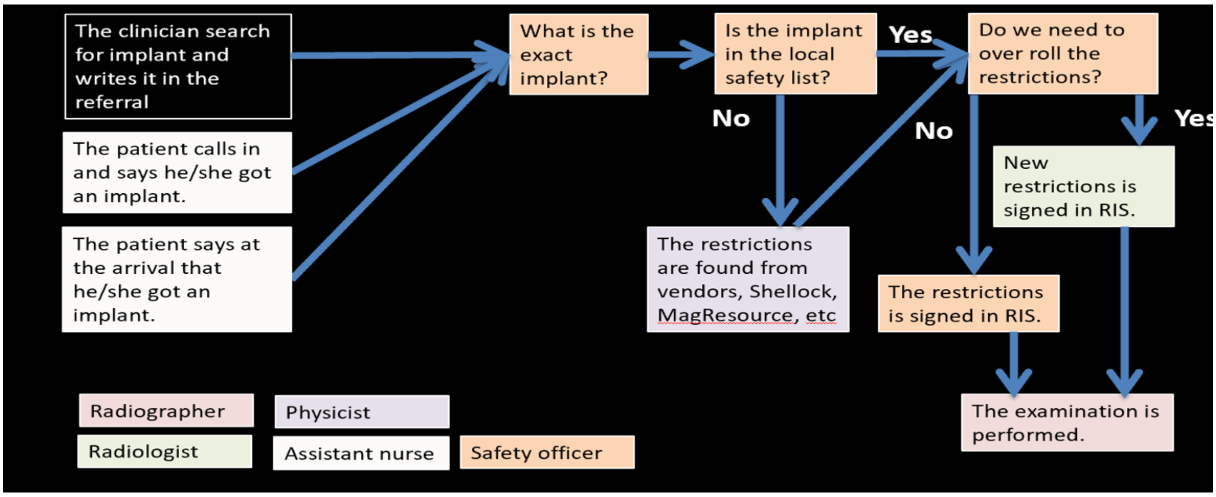

The clinician search for implant and writes it in the referral
The patient calls in and says he/she got an implant.
The patient says at the arrival that he/she got an implant.
What is the exact implant?
Is the implant in the local safety list?
Yes
No
Do we need to over roll the restrictions?
No
Yes
New restrictions is signed in RIS.
The restrictions are found from vendors, Shellock, MagResource, etc
The restrictions is signed in RIS.
The examination is performed.
Radiographer
Physicist
Radiologist
Assistant nurse
Safety officer

Figure 2: Current workflow (Kihlberg and Lundberg, 2019)

nition (NER), where the entities to be identified are words indicative of implants. We considered the possibility of fine-tuning a BERT pre-trained model on a labelled corpus of implant terms using custom labels. However, at present, this annotation endevour cannot be undertaken because it requires financial resources and a time span that are not available at the time of this publication.

We then explored the unsupervised NER solution based on BERT proposed in an article by Ajit Rajasekharan[2]. This article describes a fully-unsupervised approach to NER based on the pre-trained *bert-large-cased* (English). The approach relies on signatures indicating entities, on a morphological tagger and on BERT's Masked Language Model (MLM) head to predict candidate words for the masked positions. To put it simply, the approach combines NER and MLM using the head of the MLM to extract entities. Results seem to be promising for the NER task. However, the adaptation of this approach to the recognition of implant terms and to our domain-specific data resulted overly complex. One main hinder to this adaptation is the use of a morphological tagger. Our data is noisy and specialized and the result of a tagger on this data is certainly unreliable without a proper retraining of the tagger itself for domain and genre. Another difficult step to adapt to our task is the creation of signatures that are then handled at raw word embedding level. As the author puts it, the unsupervised NER approach works because: "BERT's raw word embeddings capture useful and separable information (distinct histogram tails with less than 0.1 % of vocabulary) about a term using other words in BERT's vocabulary. [...] The transformed versions of these embeddings output by a BERT model with MLM head are used to make predictions of masked words. The predictions also have a distinct tail. This is used to choose the context sensitive signature for a term.". First of all, the extraction of the signatures would be redundant in our case since we have already a list of implant terms (i.e the glossaries described in Section 4.1.3) automatically extracted from existing official documentation and used as search terms. Second, many of these terms are not, presumably, in the general-purpose vocabulary of the pre-trained BERT, since they are very specialized. Essentially, we ended up with the conclusion that it would be very time-consuming (if ever possible) to implement some of the steps in the fully-unsupervised BERT-NER model. The approach remains indeed inspiring and could work better or streamlined if we could re-train BERT on our domain-specific corpus, an operation that we are unable to carry out at the time of this publication.

We also explored the possibility of using a BERT model specifically fine-tuned on our corpus to predict *masked tokens* to find candidate implant terms. However, we realized that such an approach is dwarfed, because only the words in the vocabulary of the pre-trained BERT model would be suggested. If a term, e.g. "shunt", is not in that vocabulary or cannot be reconstructed using BERT tokens, it will never be "discovered" as implant term and will remain undetected.

In the experiments presented here we build on re-

[2] https://towardsdatascience.com/unsupervised-ner-using-bert-2d7af5f90b8a (published 2020, updated 2021, retrieved 2021.)

search 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 EMRs belonging to the cardiology clinic (see section 3), Kindberg (2019) evaluated 500 terms, i.e. 10 search words and their 50 closest neighbours. For the evaluation, all the terms were divided into 14 categories, and only three of these categories contained words indicative of implants. All in all, 26.2% of the 500 analysed words were considered words indicative of implants, i.e. "synonyms, semantically similar terms, abbreviations, misspelled terms" (p. 13).

For the same task on the same cardiology clinic, Nilsson et al. (2020) used Swedish BERT (see Section 4.1). The results presented in Nilsson et al. (2020) showed that "[o]ut of the 148 evaluated queries, 68 query words (46%) in their given context were considered to be clearly indicative for implants or other harmful objects. 27 query words (18%) were considered possibly indicative and 53 query words (36%) were considered non-indicative. For each query that was clearly or possibly indicative, five contextually similar words were identified which resulted in 475 additional words in given contexts. Among these 475 additional words, 83 (17,5%) words were considered as clearly indicative in their context, 105 (22%) were considered as possibly indicative and 287 (60,5%) were considered non-indicative. 40% of the 475 additional words identified with the KD-Tree queries and BERT were deemed to be possibly indicative or clearly indicative of implants or other harmful objects." (p.23-24).

It must be noticed that the results by Kindberg (2019) and by Nilsson et al. (2020) are not directly comparable between them since the evaluation methods are different. Although we learned a lot from these two previous studies, we are unable to compare our results with theirs, because in our experiments we create a model on two clinics, i.e. cardiology and neurology, rather than only on cardiology. What is more, our evaluation methods and metrics differ from those utilized by Kindberg (2019) and Nilsson et al. (2020).

3 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 neurology clinic. The 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 of record distribution in shown in Table 1). These EMRs vary greatly in length, from just a few words to hundreds of words. This data has not yet been fully anonymised, therefore we are unable to release the datasets at the time of this publication. However, we will distribute secondary linguistic data, such as automatically created wordlists on the project website.

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

Table 1: Number of words and EMRs per clinic.

4 Method: BERT

Previous methods to represent features as vectors were unable to capture the context of individual words in the 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 word 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 (Mikolov et al., 2013). 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 (Peters et al., 2018). 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 finetuned for several downstream tasks.

BERT is a powerful but complex model. According to the Occam's razor principle, simplicity must be preferred whenever possible. To comply to this principle, we carried out a few preliminary experiments on samples taken from the current dataset (cardiology + neurology) with approaches less complex than BERT, like distributional semantics based on BOW[3] and Word2Vec[4]. Results on the samples showed that BERT performed better than the others methods. A comparative study on the whole dataset (not only samples) is in preparation. These results, together with additional experiments that are still in progress, will also be available in the final project's report that will be handed in to the funding body.

In the experiments presented here, we fine-tuned BERT for focussed terminology extraction and relied on PyTorch (an open source machine learning framework[5]) (Paszke et al., 2019) and used the Huggingface transformers library for BERT (Wolf et al., 2019) available and ready to use[6].

4.1 Swedish BERT

4.1.1 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 (Malmsten et al., 2020)[7]. 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 2.

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	$11 \cdot 10^7$

Table 2: Pre-training parameters

4.1.2 Fine-Tuning the Pre-Trained Model: Phase 1

We call this tine-tuning "Phase 1" because in the near future we are going to try out different fine-tuning configurations in order to understand how to determine the optimal hyperparameters' settings for the task at hand. In Phase 1, the decisions about how to set parameters were made partly based on the original BERT paper (Devlin et al., 2019), partly on previous findings based on electronic health records notes (Li et al., 2019), partly on the observation of our current data. Hyper-parameters used for fine-tuning in this study are shown in Table 3. We relied on the Adam algorithm with default values for its hyperparameters as indicated by (Kingma and Ba, 2014). The pre-processed EMRs and the pre-trained model were fed into a Python script.

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

Table 3: Parameters used for fine-tuning

The model was fine-tuned with MLM (Masked Language Model), 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 block size was set to 64, which means that

[3] To find synonyms or semantically related words, the *textstat_simil* function of the Quanteda R package (Benoit et al., 2018) was used.

[4] Package 'word2vec, R wrapper, https://cran.r-project.org/web/packages/word2vec/word2vec.pdf

[5] https://pytorch.org/

[6] https://huggingface.co/transformers/

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

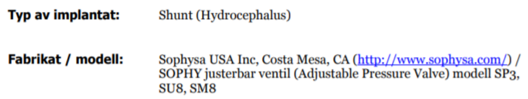
sequences with fewer than 64 tokens are padded to meet this length, and sequences with more than 64 tokens are truncated. Actually, the value of 64 is generous since according to our current calculations the average sentence length in tokens is 12. The fine-tuning took approximately 15 hours per clinic to complete using the computing resources shown in Table 4.

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

Table 4: Details of computing resources.

4.1.3 Discovering Contextually-Similar Implant Terms

We used the MRI-safety handbook (SMRlink) available at the hospital website to automatically create glossaries of implants or implant-related terms. In these experiments, we used two glossary versions, an extended version containing 753 terms that include some noise, i.e. non-implant terms, and a baseline version containing 461 terms and less noise, but also fewer terms. The extended glossary was automatically built from several sections of the documents that can be found in SMRlink. The baseline version was extracted only from the headings ”Typ av implatat” and ”Fabrikat / modell” (see Figure 3). The advantage of the extended version is the presence of potentially more implant terms. Neither of the two glossary versions was validated by domain experts, since we wanted to limit human intervention as much as possible and explore the effect of different choices.



Typ av implantat: Shunt (Hydrocephalus)

Fabrikat / modell: Sophysa USA Inc, Costa Mesa, CA (http://www.sophysa.com/) / SOPHY justerbar ventil (Adjustable Pressure Valve) modell SP3, SU8, SM8

Figure 3: The terms in the baseline glossary were extracted from the headings shown in this picture.

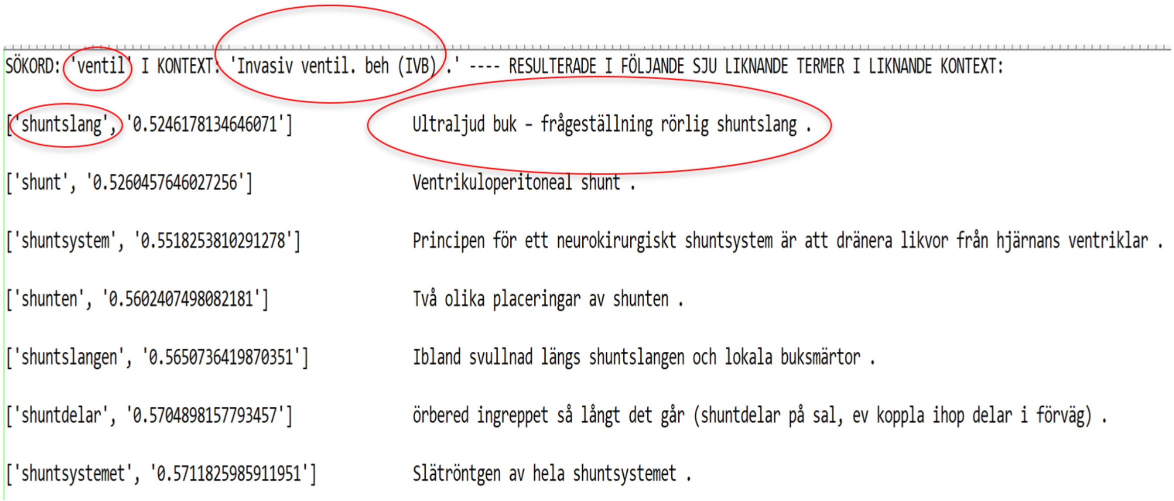
With glossary terms and the corpus, queries were created. A query is basically an example sentence containing a glossary term. Our queries are randomly chosen in the corpus. The model retrieves sentences similar to the queries and extract the term that is most similar to the glossary term that the queries exemplify (see Figure 4).

In this paper, we present the results of a BERT model evaluated using 15 queries for each glossary term. The queries were randomly chosen and used to find contextually similar sentences. This BERT model first identifies sentences in the corpus that are similar to the queries, then it extracts words in the BERT “discovered” sentences that have similar syntactic/semantic role/slot (i.e. the same “semantic role” in a broad sense) as the glossary terms that were used to build the queries. Since our corpus is sizeable, we decided that pairwise cosine similarity metric (brute force) would have been too inefficient with ordinary computing resources, and not compliant to the Green NLP paradigm (Derczynski, 2020). To build the search space we used instead the scikit-learn implementation of the KD-Tree and BallTree algorithms (Pedregosa et al., 2011), both with default distance metrics. KDTree (short for k-dimensional tree) 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[8]). While the KDTree is “a binary tree structure which recursively partitions the parameter space along the data axes, dividing it into nested orthotropic regions into which data points are filed”, BallTrees “partition[s] data in a series of nesting hyper-spheres. This makes tree construction more costly than that of the KD tree, but results in a data structure which can be very efficient on highly structured data, even in very high dimensions”. KDTree and BallTree are both memory-intensive. In order to speed up this part of the computation, the data was split into chunks. Each individual chunk 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 at the time.

4.2 Evaluation

To judge whether a term discovered using this BERT model is indicative of the presence of implants, special domain knowledge is required. In some cases, it may be obvious that a term indicates implants. In other cases, it may be less obvious due to very domain-specific sublanguage. For this reason, manual evaluation of BERT discoveries was carried out by two MRI-physicists from the

[8] https://scikit-learn.org/stable/modules/neighbors.htm



```
SÖKORD: 'ventil' I KONTEXT: 'Invasiv ventil. beh (IVB) .' ---- RESULTERADE I FÖLJANDE SJU LIKNANDE TERMER I LIKNANDE KONTEXT:

['shuntslang', '0.5246178134646071']          Ultraljud buk - frågeställning rörlig shuntslang .

['shunt', '0.5260457646027256']               Ventrikuloperitoneal shunt .

['shuntsystem', '0.5518253810291278']         Principen för ett neurokirurgiskt shuntsystem är att dränera likvor från hjärnans ventriklar .

['shunten', '0.5602407498082181']             Två olika placeringar av shunten .

['shuntslangen', '0.5650736419870351']        Ibland svullnad längs shuntslangen och lokala buksmärtor .

['shuntdelar', '0.5704898157793457']          örbered ingreppet så långt det går (shuntdelar på sal, ev koppla ihop delar i förväg) .

['shuntsystemet', '0.5711825985911951']       Slätröntgen av hela shuntsystemet .
```

Figure 4: A mock-up of the results retrieved by a query: 'Sökord' (en: search term) is a glossary term. 'I kontext' (en: in context) is a query where the search term appears. The model extract 7 sentences similar to the query and extracts terms contexually similar to the glossary term.

Radiology clinic at Linköping University Hospital, who assessed independently the terms discovered by the BERT model. For the evaluation with used the results obtained with KDTree and the extended glossary, which amount to 4636 BERT terms. More specifically, we started up with 753 glossary terms (unigrams) including noise; for each glossary term, a set of 15 queries was created (15 is an arbitrary choice); KDTree was used to search the vector space from which we extracted 7 nearest neighbours for a given query (7 is an arbitrary choice) (see Figure 4); then we merged the results for all the queries together and removed duplicates.

The two MRI-physicists received an excel file containing the list of terms to be assessed without any context, 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 judgements. Results are presented in the next section.

5 Results and Evaluation

Inter-Rater Agreement. 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 (Cohen, 1960) and Krippendorff's alpha (Krippendorff, 1980) 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" (Artstein and Poesio, 2008). 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 (Artstein and Poesio, 2008).

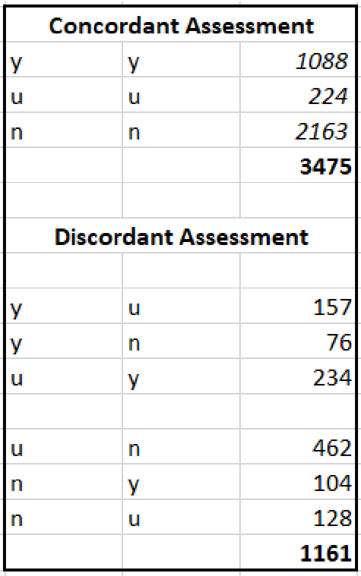
Terms	Percentage	Cohen's Kappa	Krippendorff's Alpha
4636	75%	0.575	0.573

Table 5: Inter-rater agreement on 4636 BERT terms.

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%)

Table 6: Breakdown by rater.

Tables 5 and 6 show the breakdown of the inter-rater agreement of the 4636 terms discovered by BERT. The raters agree on 3475 terms, of which **1088** were assessed to be indicative implant terms (approx. 23.5%), 2163 terms were assessed not to be indicative of implants, and for 224 terms both raters agreed on being "unsure". The raters 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 5). Out of the 1088 BERT terms indicative of implants, about 900 were not in the extended glossary and more than 1000 were not present in the baseline glossary, e.g. 'carillon-device' or 'cochlea'. Therefore these BERT terms make a useful addition to the glossaries. Out of 2163 non-indicative BERT terms, about 2000 were not in the glossaries, which suggests that the level of noise in the glossaries is relatively small.



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

Figure 5: Breakdown: concordant/discordant assessments by the two raters.

Overall, the values in Table 5 show 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 (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.

Gold Standard and Term Clusters: ***Intra-cluster Cleanliness***. The evaluateD BERT terms are the first building block of a gold standard for this task. We use this "ground truth" to assess the quality of the individual term clusters. In this context, a term cluster is a group of words semantically-related to a glossary term used to build queries (see Section 4.1.3). Examples of term clusters are shown in the Appendix.

Since we will never know the number of True Negatives and False Negatives in this task, we cannot use traditional evaluation metrics. For this reason, we used a metric that we call "term cluster cleanliness" (short **cleanliness**) to roughly assess the linguistic quality and the term relatedness within a cluster.

Cleanliness is the proportion of True Positives (TP) with respect to the numbers of terms in the cluster, i.e.:

Cleanliness= TP/(TP + FP + U + Disc + New)

where:

TP (True Positives) is the number of terms that are classified as *indicative* of implants by both annotators in the gold standard.

FP (False Positives) is the number of terms that are classified as *non-indicative* of implants by both annotators in the gold standard;

U (Unsure) is the number of terms that both annotators agree on being unsure about whether they are indicative of implants or not;

Disc (Discordant) is the number of terms in the gold standard on which the annotators disagree upon.

New is the number of terms that are not in the gold standard but are in a cluster.

This metric is simple, but handy. Additionally, numbers can be easily swapped in the formula, so that it is possible to account for the proportion of new terms (Novelty) or Undecidedness, etc. For instance:

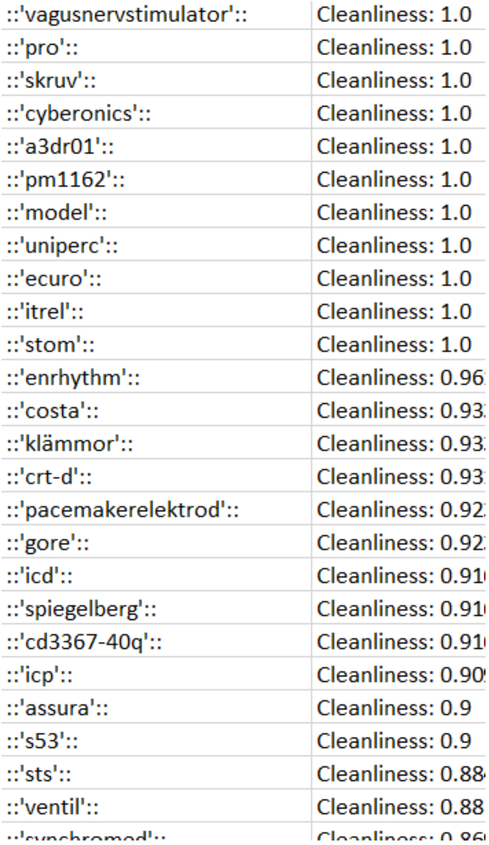
Novelty = New/(TP + FP + U + Disc + New)

Undecidedness = U/(TP + FP + U + Disc + New)

The cleanliness scores can be used to rank the term clusters and to set a threshold to trim out uninteresting terms (Figure 6 shows the top-ranked clusters returned by BallTree with extended glossary).

6 Discussion

The combination of searching the result space and the two versions of the glossary show that differing clusters are produced for the same glossary term (see the results for the glossary term 'ventil' in the Appendix, Figures A1, A2, A3 and A4). One possible way to unify these nuanced results would be to select the cluster with the highest cleanliness score for the same glossary term. For instance, for



::'vagusnervstimulator'::	Cleanliness: 1.0
::'pro'::	Cleanliness: 1.0
::'skruv'::	Cleanliness: 1.0
::'cyberonics'::	Cleanliness: 1.0
::'a3dr01'::	Cleanliness: 1.0
::'pm1162'::	Cleanliness: 1.0
::'model'::	Cleanliness: 1.0
::'uniperc'::	Cleanliness: 1.0
::'ecuro'::	Cleanliness: 1.0
::'itrel'::	Cleanliness: 1.0
::'stom'::	Cleanliness: 1.0
::'enrhythm'::	Cleanliness: 0.96
::'costa'::	Cleanliness: 0.93
::'klämmor'::	Cleanliness: 0.93
::'crt-d'::	Cleanliness: 0.93
::'pacemakerelektrod'::	Cleanliness: 0.92
::'gore'::	Cleanliness: 0.92
::'icd'::	Cleanliness: 0.91
::'spiegelberg'::	Cleanliness: 0.91
::'cd3367-40q'::	Cleanliness: 0.91
::'icp'::	Cleanliness: 0.90
::'assura'::	Cleanliness: 0.9
::'s53'::	Cleanliness: 0.9
::'sts'::	Cleanliness: 0.88
::'ventil'::	Cleanliness: 0.88
::'synchromed'::	Cleanliness: 0.86

Figure 6: Top-ranked term clusters (BallTree, extended glossary).

the term ‘ventil’, the best cluster is the one shown in Figure A2, since it has the best score.

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 to identify the kind of irrelevant words the model retrieves. Error analysis indicates that families of irrelevant words negatively affect the quality of the clusters, e.g. named entities, like *Ann-Christin* (see Figure A5 in the Appendix) and general medical terms, like *aneurysm* (see Figure A6 in the Appendix). The next step is then to filter out semantic families of words that create noise in the results. However, this operation is not straightforward since there are some apparently non-indicative words (like ‘obs’, en: attention) that helped in the discovery of implant terms because they frequently co-occur with them (see Figure A7 in the Appendix). This means that ranking the term clusters based only on their cleanliness is helpful, but it does not tell the whole story about how indicative words can be in domain-specific contexts.

7 Conclusion

In this paper, we presented results of a BERT model for focused terminology extraction. The model was devised to discover terms indicative of implants in Swedish EMRs. Although the task is challenging, manual evaluation shows that the approach is rewarding, since a solid number of indicative terms were discovered by BERT. We used these BERT discoveries assessed by domain experts to create the first building block of a gold standard that we will use to evaluate future versions of our model. We plan the following:

- annotation of the “new” terms (cyan spheres in the figures in the Appendix) by the two rater; these terms and their annotation will be appended to the current gold standard;
- the removal of named entities mentioned in the texts of the EMRs;
- the removal of general medical terms;
- the cleansing of the noise in the glossaries using the non-indicative words annotated during the creation of the gold standard;
- the conflation of the cleaned baseline and extended glossary into a single one;
- a deeper understanding of the effect of fine-tuning parameters (e.g. the effect of a smaller block size);
- a more advanced search method of the result space to overcome the fragmentation of the corpus in data parts (chunks) of 50000 tokenized sentences at the time in order to avoid the re-merging of all the results at the end of this process.

Acknowledgements

Thanks to the reviewers for insightful comments and useful suggestions. 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 (Etikprövningsmyndigheten), authorizaton nr.: 2021-00890 (to Peter Lundberg). ***Project Website***: http://www.santini.se/mri-terms/

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774.

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

Gabor Csardi and Tamas Nepusz. 2006. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

Hercules Dalianis, Martin Hassel, and Sumithra Velupillai. 2009. The Stockholm EPR corpus-characteristics and some initial findings. In *Proceedings of ISHIMR*, pages 243–249.

Leon Derczynski. 2020. Power consumption variation over activation functions. *arXiv preprint arXiv:2006.07237*.

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.

Aron Henriksson, Hans Moen, Maria Skeppstedt, Ann-Marie Eklund, Vidas Daudaravicius, and Martin Hassel. 2012. Synonym extraction of medical terms from clinical text using combinations of word space models. *Proceedings of Semantic Mining in Biomedicine (SMBM). Institute of Computational Linguistics, University of Zurich*, pages 10–17.

Ashwin Ittoo and Gosse Bouma. 2013. Term extraction from sparse, ungrammatical domain-specific documents. *Expert Systems with Applications*, 40(7):2530–2540.

Johan Kihlberg and Peter Lundberg. 2019. Improved workflow with implants gave more satisfied staff. In *SMRT 28th Annual Meeting 10-13 May 2019*.

Erik Kindberg. 2019. Word embeddings and patient records: The identification of MRI risk patients. B.sc. thesis, Linköping University.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv*, pages arXiv–1412.

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

Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability. `http://repository.upenn.edu/asc_papers/43`.

Fei Li, Yonghao Jin, Weisong Liu, Bhanu Pratap Singh Rawat, Pengshan Cai, and Hong Yu. 2019. Fine-tuning bidirectional encoder representations from transformers (BERT)–based models on large-scale electronic health record notes: An empirical study. *JMIR medical informatics*, 7(3):e14830.

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.

Anton Nilsson, Jonathan Källbäcker, Julius Monsen, Linda Nilsson, Marianne Mattila, Martin Jakobsson, and Oskar 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.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. 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)*, pages 2227–2237.

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

Elliot Schumacher and Mark Dredze. 2019. Learning unsupervised contextual representations for medical synonym discovery. *JAMIA open*, 2(4):538–546.

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.

Appendix

The graphs in this section are created with the R package Igraph[9](Csardi and Nepusz, 2006).

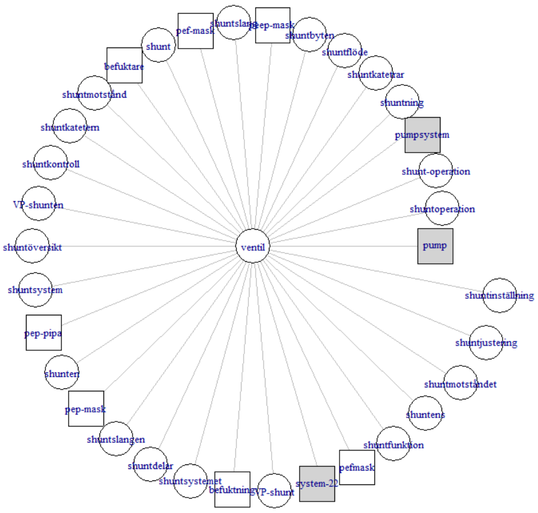


Figure A1: **KDTree, extended glossary: BERT terms related to ‘ventil’ (en: valve). Cleanliness: 0.70** *Legend*: blank circles are TPs, blank squares are FPs, grey squares are Disc. The length of the edges represents the distance of a BERT term from the glossary term.

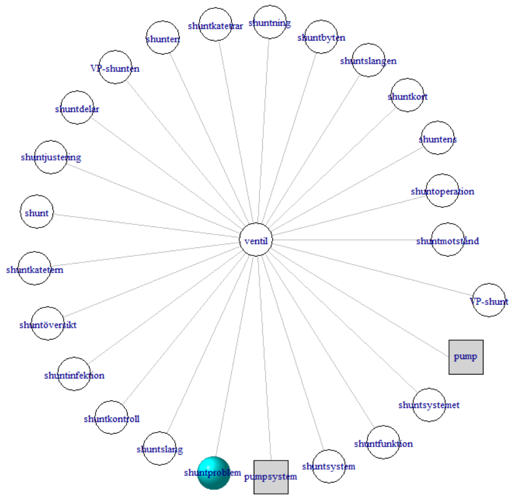


Figure A2: **BallTree, extended glossary: BERT terms related to ‘ventil’ (en: valve). Cleanliness: 0.88** *Legend*: blank circles are TPs, blank squares are FPs, grey squares are Disc, cyan spheres are words that are not in the gold standard. The length of the edges represents the distance of a BERT term from the glossary term.

[9] https://cran.r-project.org/web/packages/igraph/index.html

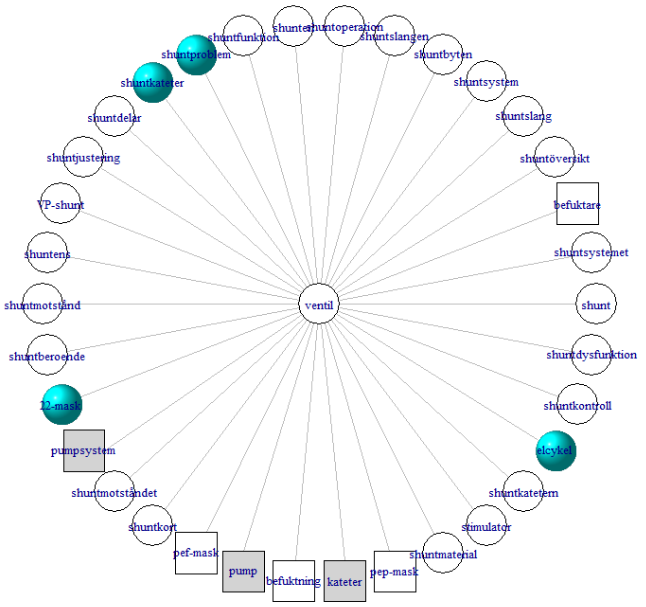


Figure A3: **KDTree, baseline glossary: BERT terms related to ‘ventil’ (en: valve). Cleanliness: 0.67**
Legend: blank circles are TPs, blank squares are FPs, grey squares are Disc, cyan spheres are words that are not in the gold standard. The length of the edges represents the distance of a BERT term from the glossary term.

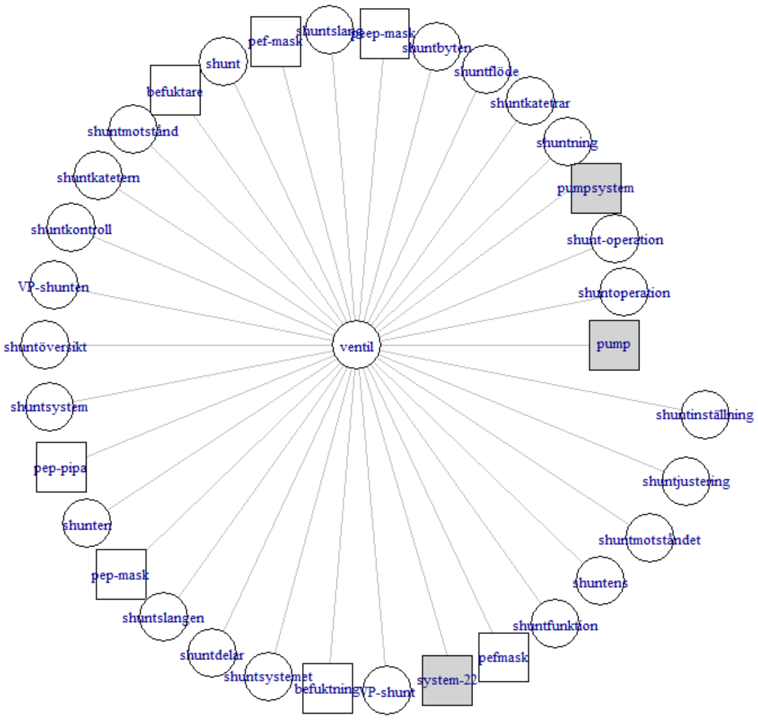


Figure A4: **BallTree, baseline glossary: BERT terms related to ‘ventil’ (en: valve). Cleanliness: 0.70**
Legend: blank circles are TPs, blank squares are FPs, grey squares are Disc. The length of the edges represents the distance of a BERT term from the glossary term.

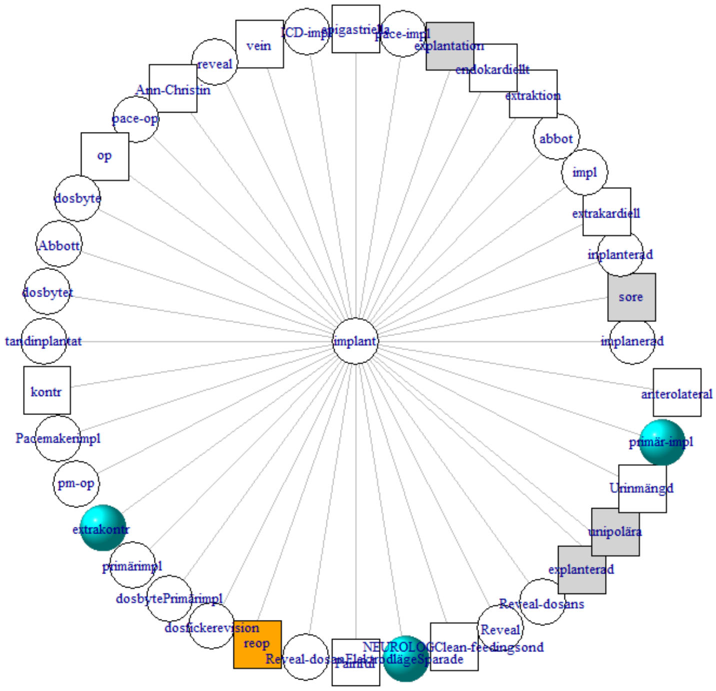


Figure A5: **BallTree, extended glossary: BERT terms related to 'implant' (in English in the glossary). Cleanliness: 0.5**

Legend: blank circles are TPs, blank squares are FPs, grey squares are Disc, orange squares are terms on which both raters are unsure about, cyan spheres are words that are not in the gold standard. The length of the edges represents the distance of a BERT term from the glossary term.

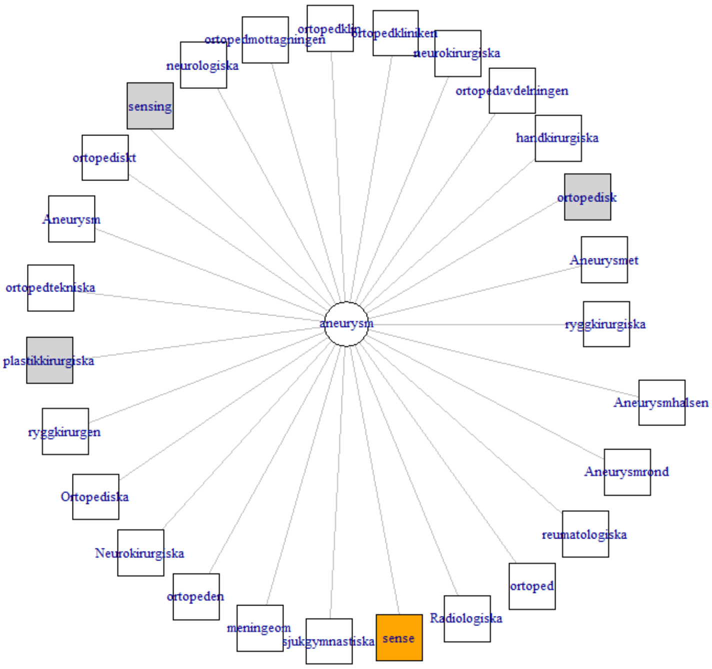


Figure A6: **BallTree, extended glossary: BERT terms related to 'aneurism'. Cleanliness: 0.0**

Legend: blank circles are TPs, blank squares are FPs, grey squares are Disc, orange squares are terms on which both raters are unsure about, cyan spheres are words that are not in the gold standard. The length of the edges represents the distance of a BERT term from the glossary term.

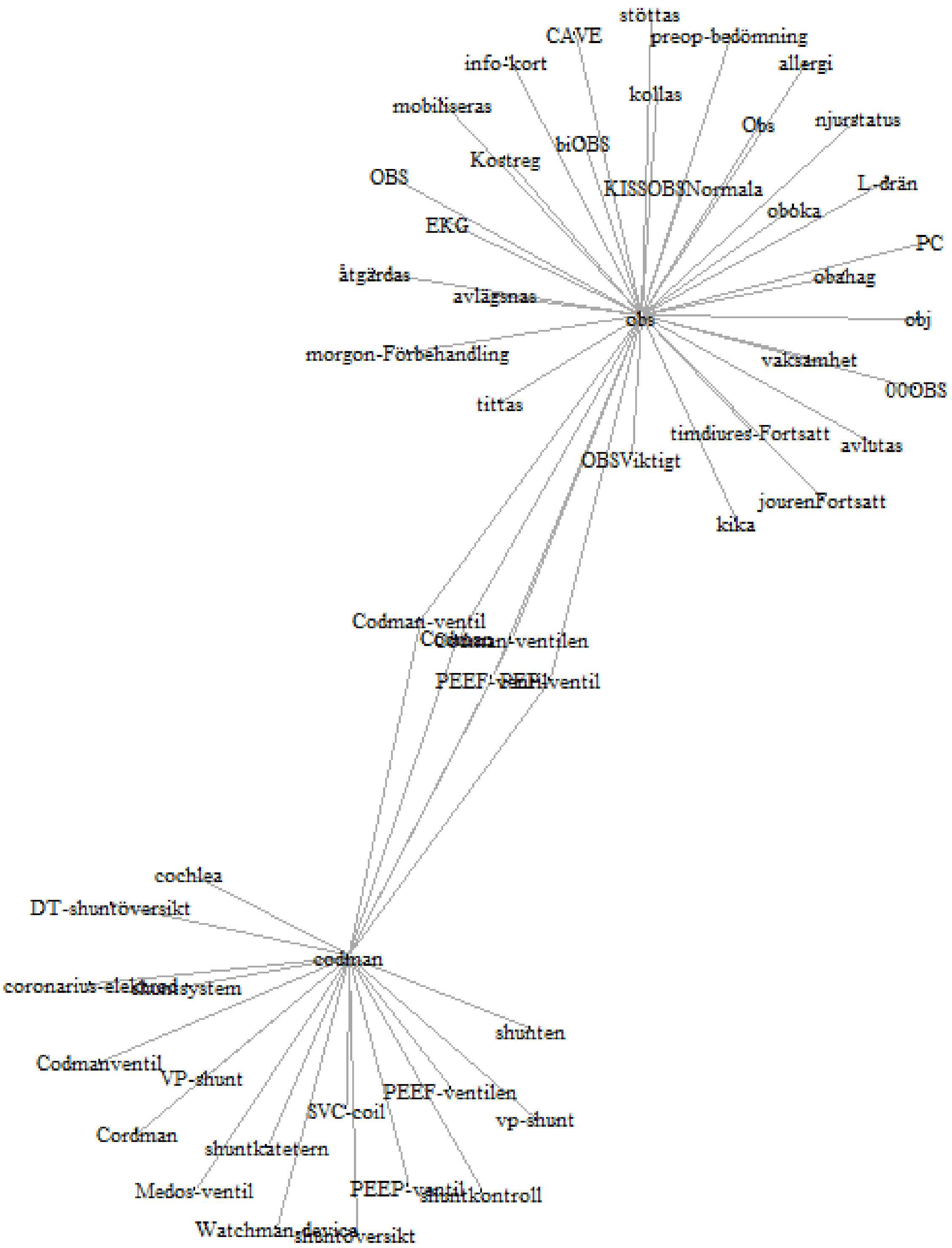


Figure A7: BallTree, extended glossary: relatedness between ‘codman’ and ‘obs