Identifying implants in patient journals using BERT and Glossary extraction

Anton Nilsson, Jonathan Källbäcker, Julius Monsen, Linda Nilsson, Marianne Mattila, Martin Jakobsson & Oskar Jerdhaf

Linköping University

729G81: Applied Cognitive Science

Examiner: Arne Jönsson

June 4, 2020

Abstract

The screening process for an MRI-scan may be resource-demanding since doctors sometimes must go through patient journal notes manually making sure patients do not have implants or other objects that may cause harm. As medical personnel are often strapped for time, journal notes may be rushed and disorganized which can result in misspellings as well as frequent abbreviations and incomplete sentences. In this study we explore the possibility of analysing patient journals to find mentioned implants or objects, that can affect MRI-scans. To achieve this, a glossary was created by means of terminology extraction from SMR-link (Safe Magnetic Resonance Database Linköping) and used for finding relevant information in provided journals. Furthermore, a BERT (Bidirectional Encoder Representations from Transformers) model was fine-tuned and used to identify terms, in certain contexts, similar to the glossary terms and the contexts in which they appeared in. The extracted words and the contextually similar terms were evaluated by domain experts that were to judge how indicative terms were for implants and other harmful objects in their given context. In total 148 glossary terms were evaluated as well as an additional 475 contextually similar terms. We found that 64% of the glossary terms were considered possibly or clearly indicative whereas 40% of the contextually similar terms were considered possibly or clearly indicative. We also found that contextually similar terms were more likely to be evaluated as clearly indicative if the glossary term they were based on, were evaluated as clearly indicative. These results show great promise for using these methods in this domain specific task. However, we did not utilize all the provided data in this study and the end results could have significantly benefitted from further refining of the glossary and the contexts provided by keyword matching. Some inter-rater reliability in evaluating the results would also have been preferred. Hence, future work is needed to explore these and other aspects further.

1 Introduction

MRI-scanning is a powerful tool that can help us explore the human body without invasive procedures. However, implants such as pacemakers can injure the body or cause discomfort during an MRI-scan, while others may cause damage to the expensive equipment. As a patient is remitted for an MRI-scan, they are required to fill out a form regarding any implants or other metal or electronic objects in their bodies. If there are any doubts about the patient's testimony regarding implants, the medical staff will manually go through their journals to make sure that the scan can be conducted safely. As medical personnel are often strapped for time, journal notes may be rushed and disorganized which can result in misspellings as well as frequent abbreviations and incomplete sentences (Allvin et al., 2011). Today the process of going through the journals is done manually, which is a waste of valuable time and resources. Luckily, this time-consuming task could be automated with the aid of computers and language technology applications.

This could be done by creating a wordlist of relevant terminology and attempt to match the words to implants in the journals. However, because of the rushed journal notes, there are words in the journals that might not explicitly indicate harmful implants while still implying their existence. As a result, a word-list matching could fail to identify these implants. To be able to address this problem, one can use language modelling. An example of this is BERT (Bidirectional Encoder Representations from Transformers), which is a language technology model created by Google that can, by analysing sentences, understand the contextual relations between words in a text. By the employment of both terminology extraction for the creation of a glossary along with a BERT-model tuned to medical data, the usefulness of these methods can be examined as a first step to create an automated solution for this problem.

1.1 Purpose

The purpose of the project is to examine the possibility of analysing patient journals to find potential implants or objects that can affect MRI-scans. By doing this the hope is to lay the groundwork for future studies on this and similar subjects.

1.2 Research Goals

Given the background, this study had three goals. This study aimed to:

- Examine how a glossary of key-terms could be used for identifying mentioned cases of implants in patient journals.
- Examine if BERT can be used to identify abbreviations, synonyms and semantically similar terms with spellings errors in patient journal data.
- Examine the possibility of combining BERT and keyword matching based on a glossary of extracted terms.

2 Background

The background will present information regarding neural networks and language modelling. This includes information about word embeddings and sentence similarities and a description of the BERT-model. It also handles previous applied research, information on the K-nearest neighbour algorithm and a description of terminology extraction.

2.1 Neural networks and Language modelling

Language modelling is the task of creating representations of language in ways in which machines can understand how language works. In recent years an approach for modelling language has gained popularity, namely that of neural networks. Artificial neural networks (ANN) were originally created to mimic the way the human brain works. Typically, and ANN consists of an input layer, a number of hidden layers, which can vary in number, and at last, an output layer.

The neurons in the different layers receive and transform data using an activation function which is a mathematical equation that determines if a neuron should fire a signal or not (Winters-Miner et al., 2015).

2.1.1 Word embeddings and sentence similarity

Word embeddings are used to represent words in neural networks. A word embedding is a representation of a word as a vector of numbers, in possibly hundreds of dimensions, in a predefined vector space (Wang et al., 2019). Word embeddings can be used to measure similarities between words. An example of one such measure is cosine similarity. It measures the cosine of the angle between the two vectors, which means that it is measuring if the two vectors are pointing in the same direction in a vector space. (Wang et al., 2019).

2.1.2 BERT

The recent Google developed method BERT (Bidirectional Encoder Representations from Transformers) has contributed to significant progress in natural language processing (NLP). This is because it can pre-train deep bidirectional representations of language on unlabelled data, and further be fine-tuned with minimal data and effort for specific tasks (Devlin, Chang, & Toutanova, 2019). BERT is made up of several transformer encoders and does hence take advantage of an attention mechanisms that enable the model to learn contextual relations between words in a text. The bidirectionality in BERT comes from the fact that it reads entire sequences of words at once and learns representations from both left and right contexts of words. Furthermore, BERT is useful for extracting high-quality language features from text (McCormick & Ryan, 2019). These features can be token embedding vectors which can be combined in ways to generate word and sentence level embeddings.

2.2 Previous applied research

Earlier models such as Word2Vec has been evaluated both qualitative and quantitative in the task of finding semantic relatedness and similarity of biomedical terms (Zhu, Yan, & Wang, 2017). Similarly, one study investigated and evaluated Word2Vec´s usefulness when it came to find semantically related terms in patient journals in the context of MRI examinations (Kindberg, 2019). This study was also conducted at Linköping University Hospital radiology clinic on a part of the data used in the current study. More recent studies have taken advantage of the recent developments in NLP by applying contextualized deep learning representation models, such as ELMo (Peters, et al., 2018) and BERT (Devlin, Chang, & Toutanova, 2019). For example, Schumacher & Dredze (2019) used these models to build representations of and identify synonymous words or phrases in medical data by the contextual features of words.

2.3 KNN and KD-Trees

The K-nearest neighbour algorithm is an algorithm that works by discriminating data, such as word embeddings, based on feature similarities. (Pedregosa, et al., 2020). One specific implementation of the algorithm is called the K-dimensions tree, or KD-Tree (Bentley, 1975). The advantage of this variation of the algorithm is that it saves computational cost. It makes the comparison based on an area within the median of the attributes in the training data. This results in the nearest neighbour comparison being based on the samples that are found within that area.

2.4 Terminology Extraction

Terminology extraction (or key-word extraction) is a well-researched area of natural language processing that concerns automatically selecting terms from texts (Woo, Kim, & Lee, 2020). Statistical methods utilize properties like word frequencies in order to build versatile models, while rule-based methods require manually created rules. The hybrid approach is based on a combination of the statistical and rule-based methods and attempts to improve performance through sophisticated combinations of word frequencies and rules. Possible applications for terminology extraction include, for example, text labelling, text summarization, and vocabulary creation. Vocabulary or glossary extraction is the process of extracting relevant terms from a larger corpus in order to create a vocabulary, which can further be used to analyse new texts (Park, Byrd, & Boguraev, 2002).

3 Methods

The methods chapter will present the data in the project. It also presents the proceedings of the term extraction, the model fine-tuning and how the comparison of the contextual word embeddings was made. Furthermore, it describes how medically trained personnel evaluated the results of the BERT-model.

3.1 Material and Data

The patient journals used in the study are from two different clinics at Linköping University Hospital, the neurology-clinic and the cardiology-clinic. Using an anonymous ID of a patient all journals that contain information about that individual can be gathered and examined. SMR-Link (Safe Magnetic Resonance Database Linköping) is a safety manual used by medical professionals in Region Östergötland. The database contains information about the most common implants that may be found in patients' bodies. Implants described in SMR-Link are divided into three different safety levels: dangerous, conditional, or safe.

3.2 Term extraction

The data extracted was taken from the SMR-database using a "scraper"-python script. In simple terms the scraper opens all the webpages from the hyperlinks available on SMR-Link and extracts specific regions on each page along with any additional links it can open. From the scraped data an initial list of terms was created containing all the terms.

3.3 Model fine-tuning

The pre-trained BERT model used in this paper ("bert-base-swedish-cased") had a vocabulary of 50325 words and was released by The National

Library of Sweden (Malmsten, 2020). To improve the models understanding of this very domain specific language we added another 500 of the most common words in the journals, that were not already in the vocabulary, to the pretrained BERT-models vocabulary. To fine-tune the pretrained model from the National Library of Sweden we used the PyTorch implementation of BERT, developed by Hugging Face (2020). Due to constraints like immobile journal data and the computer lacking sufficient GPU´s, we were forced to make certain trade-offs. The model was for this reason trained on approximately a third of the journal data.

3.4 Contextual word embeddings comparison

The produced keyword list from the SMR-link, contained about 1250 words. Some of the words were filtered out and out of the remaining words, 548 words were present in the journal data. With the 548 keywords and the subset containing those keywords, a total of 1034 queries were created. The queries were subsequently used to find words with similar contextual features in the other subset which did not contain sentences with any keywords. This was done using the done using the scikit-learn implementation of the KD-Tree algorithm.

3.5 Evaluation

To judge whether a term in a given context is relevant and indicative for implants or other objects potentially harmful during an MRI-scanning, special domain knowledge may be needed. To evaluate the results gained from the queries, two MRI-physicians from the Radiology clinic at Linköping University Hospital were therefore asked to judge the results. They were to rate their judgement on a three-degree scale. 1 corresponded to the term not being indicative, 2 to the term being potentially indicative and 3 to the term being clearly indicative for implants or other harmful objects.

4 Results

Out 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.

An example of a result from a query can be found in Table 1. In this example the query word was “medtronic” which is a manufacturer of pacemakers among other medical products. This word appeared in a certain context provided by the keyword matching. The following contextually similar words were found with BERT by querying the KD-Tree for the 5 nearest neighbours. (See table below)

word	rating
pm-ddd	3
106	3
framgångsrik	3
höger	3
med	3

All these neighbours were judged to be clearly indicative of implants or other harmful objects, although a few of them may have been erroneously evaluated, for example the word “framgångsrik” (successful). On the other hand, the word “med” seems to have been correctly evaluated since it was mentioned in its context as “med sensia” where sensia is a specific pacemaker model manufactured by Medtronic. In other words, “med” most likely was short for Medtronic.

In total 83 of the suggested nearest neighbour words were clearly indicative. 72 of these were found based on queries which also clearly indicated implants or other harmful objects. This means that 87% of all new words found with BERT that were evaluated as clearly indicate came from queries which were also rated as clearly indicative. Furthermore, out of the 68 queries which were judged to be clearly indicative, 49 queries (72%) had at least one neighbour that was possibly or clearly indicative. This also means that for 19 of the clearly indicative queries, only non-indicative terms were found by the contextual embedding comparison. Additionally, a Pearson correlation coefficient was calculated between the ranking of the 5 nearest neighbours extracted with BERT and the judged relevance of the suggested words. The purpose of this was to see if the proximity of the neighbours to the query word had influence on the indicability of the terms. We found no significant correlation between the ranking of the nearest neighbours extracted with BERT and the judged relevance of the suggested words, $r(3) = -0.572$, $p = 0.314$. This suggests that the similarity ranking has no relevance.

5 Analysis and discussion

The fact that 46% and 18% of the evaluated queries were considered possibly or clearly indicative, respectively, indicates on the usefulness of matching words from a produced glossary with relevant terminology with journal data. However, the randomly sampled matches were only a small sample from a much larger variety. This means that it is hard to be sure if the glossary is in fact useful or not, but the current findings would indicate that they are. The glossary could also be improved further, as of the time that tests were run the filtration and refining was not completed. However, it is still hard to judge whether a word is relevant without knowing the context in which it appears. Further refining would require several experts to evaluate the results from the BERT model and the glossary so that inter-rater reliability can be recorded and considered.

Moving on to the 475 contextually similar terms gained from the comparison of BERT embeddings with the KD-Tree, 17,5% were clear indications and 22% were somewhat indicative of implants or other harmful objects. This indicates that BERT can be used to find important terms which are contextually similar to those from the glossary. However, having a more refined glossary could have led to more relevant queries and thus better results from the contextually similar terms.

Another factor worth considering regarding the source of noise is that only 600.000 out of 6 million sentences were used to build the KD-Tree from which contextually similar words were found. Providing an example of a query and the five most contextually similar terms gained from querying the KD-Tree showed great potential, on the one hand, for identifying abbreviations. On the other hand, the rating process can be questioned, as each term was only evaluated by a single assessor, introducing a higher risk for human error. This in conjunction with the high frequency of noisy terms. To prevent this in future studies, one could preferably use more expert evaluators and make sure that the same material is evaluated by the different evaluators. Furthermore, there could have been more contextually similar terms outside of the top five nearest neighbours that would have been more similar than some of the closest neighbours. However, the reason for choosing the top five most similar was to make the evaluation work more manageable.

There are some points to discuss regarding the methods. The glossary extraction was made difficult by the fact that few studies have worked with the same kind of data before. Further, during the filtering process we did most of the sorting manually, which could have resulted in important words being removed. There are also some aspects of the fine-tuning. As mentioned, certain trade-offs had to be made which could have affected the results. Also, we had calculated the top 500 most common words in the journal data, not already existing in the vocabulary, before cutting it down to a third. This means that we cannot be certain that these

500 words were also the most common in our new downsized data. Some aspects of the data pre-processing may have been suboptimal in some senses, since sentences with an abbreviation including a dot was treated as the end of the sentence. Some sentences not starting with a capital letter also caused problems. With regards to the embeddings, we choose to create our token vectors by taking the sum of the four last hidden layers. It is possible that some of the other recommended ways of combining layer vectors would have yielded better results for this specific task. But because of memory constraints this was the best option.

6 Conclusion and future work

Based on our findings it seems plausible that a glossary alone and combined with a language model application such as BERT can be used to simplify the process of investigating patient journals for implants. We have shown that BERT can be used to identify abbreviations, synonyms, and semantically similar terms as well as to find new terms indicative of implants expanding from a glossary. Out of the 83 new words found by BERT that were rated as clearly indicative, 72 of those were based upon queries an implant. This indicates that a good glossary is essential for finding synonyms or other new relevant terms to add to the glossary. While our current results are far from ideal, the study was only a pilot study, and with more resources it is our belief that BERT could be used for this purpose. With that said, it is also imperative to underline the importance of being thorough and precise when constructing the BERT-model and the glossary. We believe that we would have gotten a more interesting result if we had used the entire data to fine-tune the BERT model and to build the KD-Tree. Further on, expert evaluation of the glossary to indicate which terms are relevant and which terms can be removed is important to remove as much noise from the glossary as possible. Future studies should keep these two points in mind.

References

Allvin, H., Carlsson, E., Dalianis, H., Danielsson-Ojala, R., Daudaravicius, V., & Hassel, M. (2011). Characteristics if finnish and swedish intensive care nursing narratives: A comperative analysis to support the develop. *JOurnal of biomedical semantics*, 1-11.

Bentley, J. L. (September 1975). Multidimensional Binary Search Trees Used for Associative Searching. *Communications of the ACM*, ss. 509-517.

Devlin, J., Chang, M. L., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*, arXiv:1810.04805v2 [cs.CL] 24 May 2019.

Hofmann, J. M., BIemann, C., & Remus, S. (2017). Benchmarking n-grams, topic models and recurrent neural networks by close sompletions, EEGs and eye movements. i B. Sharp, F. Sèdes, & W. Lubaszewski, *Cognitive approach to natural language processing* (ss. 197-215).

HuggingFace. (2020). *Transformers: State-of-the-art Natural Language Processing for PyTorch and TensorFlow 2.0*. Retrieved from GitHub: https://github.com/huggingface/transformers, 2020-05-20

Kindberg, E. (2019). *Word Embeddings and Patient Records: The Identification of MRI Risk Patients.* Linköping: Linköping University.

Malmsten, M. (2020). *National Library of Sweden - Swedish BERT Models*. Retrieved from Github: https://github.com/Kungbib/swedish-bert-models, 2020-05-20

McCormick, C., & Ryan, N. (2019, May 14). *BERT Word Embeddings Tutorial*. Retrieved from https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/

Park, Y., Byrd, R. J., & Boguraev, B. K. (2002). *Automatic glossary extraction: Beyond terminology identification.* the 19th international conference on computational linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (den 21 5 2020). *1.6. Nearest Neighbors*. Hämtat från https://scikit-learn.org/: https://scikit-learn.org/stable/modules/neighbors.html

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv*, arXiv:1802.05365v2 [cs.CL] 22 Mar 2018.

Schumacher, E., & Dredze, M. (2019). Learning unsupervised contextual representations for medical synonym discovery. *JAMIA Open, Volume 2, Issue 4, December*, 538–546, https://doi.org/10.1093/jamiaopen/ooz057.

Wang, B., Wang, A., Chen, F., Wang, Y., & Kuo, C. (2019). Evaluating word embedding models: MEthods and experimental results. *APSIPA Transactions on signal and information processing*.

Winters-Miner, L. A., Bolding, P. S., Hill, T., Nisbet, B., Goldstein, M., HIlbe, J. M., . . . Stout, D. (2015). Prediction in medicine - The data mining algorithms of predictive analytics. i L. A. Winters-Miner, P. S. Bolding, J. M. Hilbe, M. Goldstein, T. Hill, R. Nisbet, . . . G. D. Miner, *Practical predictive analytics and decisioning systems for medicin* (ss. 239-259). Elsevier.

Woo, H., Kim, J., & Lee, W. (April 2020). Development of curriculum design support system based on word embedding and terminology extraction. *Electronics*.

Zhu, Y., Yan, E., & Wang, F. (2017). Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. *BMC Med Inform Decis Mak 17, 95*, https://doi.org/10.1186/s12911-017-0498-1.