Linköping University | Department of Computer and Information Science
Bachelor's thesis, 18 ECTS | Cognitive Science
2022 | LIU-IDA/KOGVET-G--22/003--SE

Towards terminology-based keyword extraction

Cornelius Krassow

Supervisor : Leelo Keevalik
Examiner : Agneta Gulz

External supervisor : Marina Santini

Linköpings universitet
SE–581 83 Linköping
+46 13 28 10 00 , www.liu.se

Upphovsrätt

Copyright

Abstract

The digitization of information has provided an overflow of data in many areas of society, including the clinical sector. However, confidentiality issues concerning the privacy of both clinicians and patients have hampered research into how to best deal with this kind of “clinical“ data. An example of clinical data which can be found in abundance are Electronic Medical Records, or EMR for short. EMRs contain information about a patient’s medical history, such as summarizes of earlier visits, prescribed medications and more. These EMRs can be quite extensive and reading them in full can be time-consuming, especially when considering the often hectic nature of hospital work. Giving clinicians the ability to gain insight into what information is of importance when dealing with extensive EMRs might be very useful. Keyword extraction are methods developed in the field of language technology that aim to automatically extract the most important terms or phrases from a text. Applying these methods on EMR data successfully could help provide the clinicians with a helping hand when short on time. Clinical data are very domain-specific however, requiring different kinds of expert knowledge depending on what field of medicine is being investigated. Due to the scarcity of research on not only clinical keyword extractions but clinical data as a whole, foundational groundwork in how to best deal with the domain-specific demands of a clinical keyword extractor need to be laid. By exploring how the two unsupervised approaches YAKE! and KeyBERT deal with the domain-specific task of implant-focused keyword extraction, the limitations of clinical keyword extraction are tested. Furthermore, the performance of a general BERT model in comparison to a model finetuned on domain-specific data is investigated. Finally, an attempt is made to create a domain-specific set of gold-standard keywords by combining unsupervised approaches to keyword extraction is made. The results show that unsupervised approaches perform poorly when dealing with domain-specific tasks that do not have a clear correlation to the main domain of the data. Finetuned BERT models seem to perform almost as well as a general model when tasked with implant-focused keyword extraction, although further research is needed. Finally, the use of unsupervised approaches in conjunction with manual evaluations provided by domain experts show some promise.

Acknowledgements

Special thanks to:

- Leelo Keevalik - Thesis supervisor
- Marina Santini - Project supervisor
- Linköping University Hospital - Data provider
- Peter Lundberg - Domain expert and annotator
- Tomas Bjerner - Domain expert and annotator
- Yosef Al-Abasse - Domain expert and annotator

Contents

List of Tables

1. Introduction

The development of information technology has had an effect on most areas in our society, including healthcare. According to Sun et al. (2018) electronic medical records (EMR), which medical staff use to record text, charts, graphics and data generated by hospital information systems, are now as customary as anything within the healthcare systems. With this comes a wealth of opportunities as various sources of clinical information are becoming readily available for large-scale analysis.

Although large-scale analysis of medical data in the form of EMRs have been attempted in different areas such as psychiatry (Perlis et al., 2012) and genetics (Roden et al., 2008), there is no real consensus on how data of this kind ought to be handled or what kind of technical procedures are the most suitable. By default the analysis of large texts makes use of techniques developed within the field of language technology to automate the process. As EMR data can be extensive and difficult to read, finding methods to simplify and automate the extraction of important information would be useful not only for large-scale analysis but also the medical staff themselves. Giving the medical staff quick access to the most salient information contained within the patient's medical history would allow for the patient to receive adequate treatment within a smaller time frame and thus speed up the recovery.

To lend a helping hand in order to target these issues, this thesis will explore different applications of keyword extraction; a language technology method that automatically identifies terms that represent the most relevant information inside a document (Beliga et al., 2014). Keyword extraction on clinical data, or simply "clinical keyword extraction", is to a large extent an unexplored area. The purpose of this thesis is to help build an initial outline and understanding for some of the difficulties that arise from clinical keyword extraction. Due to the absence of gold-standard keywords for the dataset used in this thesis, unsupervised approaches to keyword extraction will be explored. Their abilities to handle tasks that involve multiple medical disciplines will be tested via a domain-specific task involving the retrieval of implant-focused keywords. The prowess of BERT (Bidirectional Encoder Representations from Transformers), a language model that has shown great promise in basically all areas of natural language processing (Devlin et al., 2018), will be explored in detail via the domain-specific task. Additionally, an attempt to establish a task-specific gold-standard set of keywords for the dataset will be made by utilizing the knowledge of domain experts and the unsupervised approaches.

Research questions

- What are the limitations of unsupervised approaches when handling domain-specific tasks such as implant-focused keyword extraction from Swedish EMRs
- Can a general language transformer model, such as Swedish BERT, be applied for such domain-specific tasks?
- How can a combination of unsupervised approaches to keyword extraction be used to create a task-specific gold-standard of keywords?

2. Theory

2.1 Keyword extraction

Turney (2000) describes keywords as important phrases within documents that can be used to describe the document at hand. These keywords can be used in a multitude of different natural language processing (NLP) applications. They may for example act as a tool in information retrieval applications by functioning as a filter to sort out documents which may or may not be relevant to the task at hand. Since the quantity of readily available documents has grown to an excessive amount it has become more and more important to be able to summarize and describe documents efficiently. This is where automatic keyword extractions comes in handy. Instead of having humans read through each and every document and deciding on keywords that best describe the text at hand we can use keyword extraction methods developed within machine learning disciplines to do the same job more rapidly (Firoozeh et al., 2020).

The automatic keyword generation process can be divided into two different phases; keyword assignment and keyword extraction (Siddiqi & Sharan, 2015). During the keyword assignment phase, an assembly of words and phrases are selected as candidates for keywords. In the keyword extraction phase these candidates then get ranked in some way to allow the keyword extractor to select the best candidates as keywords for the document (Beliga et al., 2015). Different methods have been developed that not only differ in their approach to the two phases but also in what type of data they require and how well they handle language differences. Broadly these methods can be categorized into four different categories: statistical approaches, linguistic approaches, machine learning approaches and other approaches (Han & Kamber, 2006). Statistical approaches to keyword extraction use simple methods such as counting word frequencies and n-grams to determine candidate keywords for a document. This approach requires no training data and is both language and domain-independant. Linguistic approaches use different linguistic properties of the words, sentences and documents to determine the candidate keywords. Commonly examined properties include lexical, syntactic and semantic analysis. This type of approach is however quite demanding and introduce complex NLP problems. Machine learning approaches come in two types; supervised and unsupervised. Supervised learning keyword extraction models require a set of gold-standard keywords that the model can use while training. Gold-standard keywords are used as a reference for machine learning models during training to ensure that the final result of the model is similar to that of the gold-standard. Creating a gold-standard of keywords for a dataset requires manual annotation, which is both tedious and time-consuming. Additionally, if the dataset stems from a domain-specific area such as medical data, a domain-expert needs to be the one manually annotating the documents. Unsupervised learning models require no gold-standard data during training and can thus circumvent some of the issues of the supervised learning models. However, the unsupervised learning models do introduce problems such as increases in complexity and reduction of accuracy. The other approaches may also incorporate heuristic knowledge such as positioning of words, length of the document and text formatting information (Beliga et al., 2015). Which keyword extraction method best suited for the task at hand depends on multiple different variables. As medical data in the form of EMRs are highly unlikely to have been manually annotated before the keyword extraction is to be performed, a method that does not require any type of gold-standard keywords would be preferred. Additionally, to avoid error pertaining to the domain-specific nature of the medical data, a domain-independent approach is required. Finally, it would be profitable if the method would be applicable to multiple languages.

2.1.1 YAKE! (Yet Another Keyword Extractor!)

YAKE! is an automatic keyword extractor that relies on statistical features of a text to select the most important keywords. The method does not need to be trained on any particular set of documents nor does it depend on the language or the domain of the text. YAKE! devised a set of five features to establish a score of how likely each term in the text is to being a keyword. The features include the casing of the term, its positioning in the text relative to other terms, the frequency of the term, how many different contexts the term is used in and finally in how many different types of sentences the term occurs (Mangaravite et al., 2020).

The YAKE! algorithm works as follows: given a text, the algorithm divides it into sentences using a rule-based sentence segmenter called *segtok* (Leitner, 2021). The sentence is then divided into chunks if any punctuation are found, then tokens using the *web_ tokenizer* module of the segtok segmenter. Each token is then converted into lowercase, before being annotated with tag delimiters. The tag delimiters used by the algorithm include digits, unparsable content (terms that are formed by at least two punctuation marks, terms that include no digits or letters, terms formed by a combination of letters and digits and terms formed by more than one digit or alpha character), acronyms, uppercase (words that begin with an uppercase character) and parsable content (the rest). Next the algorithm iterates through the chunks and sentences created in the previous step to gather statistical data which provides each term with a certain score. The statistical data include the five aforementioned features. Then, using a sliding window approach, the algorithm looks at sequences of terms ranging from 1-gram to N-grams. Based on a few conditions such as the tags of each token in the sequence and the occurrence of common, semantically irrelevant words known as stopwords the algorithm then extracts potential candidate keywords in the form of N-grams. In the next step of the algorithm each candidate keyword is given a score based on the term scores that were calculated previously.

2.2 BERT

BERT is a language model developed by Google which makes use of an attention mechanism called Transformers to learn contextual relations between words in a text. When presented with a text input, the Transformer encoder reads the entire sequence of words at once. This bidirectional encoding allows the model to learn a words context based on its entire surroundings, which sets it apart from most models that apply a directional approach where the sequence of words get read either from left to right or the other way around.

During training, the BERT model makes use of a training strategy called the Masked Learning Model (MLM). When a language model is trained, a prediction goal has to be set. Most previous models made us of directional prediction goals such as predict the next word in the sequence “*He broke his left BLANK*“. This directional approach however limits the context learning of the language model. The MLM strategy instead replace 15 % of the words in a sequence with a “[MASK]“ token. The BERT model will then attempt to predict what word could have originally been found in that slot based on the context of the non-masked words.

Next the BERT model makes use of a strategy called “Next Sentence Prediction“ (NSP). During training, the BERT model will receive pairs of sentences as input. 50 % of the time the sentences will be part of a sequence in the text, and 50 % of the time the second sentence in the pair will be a random sentence from somewhere else in the corpus. The task of the BERT model is then to predict whether or not these pairs of sentences are part of an existent sequence in the corpus or not (Devlin et al., 2018). This additionally enhances the context learning of the BERT model.

One of the main strengths with BERT is its efficiency. BERT models that have been pre-trained on large datasets can be fine-tuned for specialized tasks. Fine-tuning essentially refers to the procedure of re-training a pre-trained language model on some new data. Instead of having to create and train an entirely new model, which can be very time-consuming, a general BERT model which has been pre-trained on a large corpus can be specialized for the task at hand. This can be useful when building BERT models that deal with domain specific data (Devlin et al., 2018).

2.2.1 Pretraining and models

This study makes use of a pretrained BERT model called "bert-base-swedish-cased" developed by KBLab at the National Library of Sweden. The model was trained on a collection of texts from digitized newspapers, official government reports, the Swedish wikipedia, legal e-deposits and social media. The variety of sources from which the texts were collected was deliberately introduced to provide a BERT model that could account for the general aspects of the Swedish language. The texts in total contained about 260 million sentences and 3500 million tokens, amounting to a total training size of about 18 gigabytes (Malmsten et al., 2020). Additionally another instance of the same pre-trained model which was fine-tuned on EMR data was used. The finetuning of this model were based on parameters used in the original BERT paper, observations made about the data (such as average sentence and word length), hardware limitations and previous research on EMRs (Jerdhaf, 2021).

2.2.2 KeyBERT

KeyBERT is a keyword extraction technique that falls under the category of the unsupervised learning approach. The method involves transforming the data into a numerical representation in the shape of vectors by utilizing pre-trained BERT models. This numerical representation of the data is usually referred to as *embeddings*. BERT is a deep-learning model that has shown great results for both similarity- and paraphrasing tasks. KeyBERT firstly utilizes the BERT models to extract an embedding that represents the entire document, then extracts word embeddings for N-gram words/phrases in the provided document. In the third and final step KeyBERT makes use of cosine similarity, a measure of similarity between two sequences of numbers, to find the word embeddings that are most similar to that of the document. The intuition then is to essentially find the words or phrases inside the text that are most similar to the document as a whole. Additionally, an embedding based on a list of terms provided by the user called "seeded terms" can be used in conjunction with the document embedding to allow for a more specialized selection of keywords (Grootendorst, 2020).

2.3 Clinical keyword extractions and the problems with unstructured medical data

Applications of different NLP techniques on medical data has seen a surge of prevalence in the past few years. Most often the medical data that is used for these applications involve EMRs written by clinicians in a free-text format. These EMRs can include information about the patients medical history, discharge summaries, family history, progress notes and so on (Tang et al., 2019). Depending on the patient the EMR can be long and tedious to read. Oftentimes clinicians work with multiple patients at once and may not have the time to fully decipher what information is of importance to the patients current situation. Being able to filter through the different notes and only read what is of utmost importance could be very beneficial both to the clinician and the patient. The patient will be provided with quicker and more accurate care while the clinician is able to spend more time treating his or her patients. A clinical keyword extractor would be able to aid clinicians in their work by describing the notes with a few keywords and phrases that give an overview of what type of information may be found inside.

Clinical keyword extractions may also be beneficial for the large-scale analysis of medical data. Keywords are essential in many aspects of information retrieval such as text summarization and web searches (Bracewell et al., 2005). Being able to accurately portray the information that can be found within certain unstructured forms of medical data will help aid other NLP applications in their efforts to only analyze data that fit their specific framework. This will save whoever is trying to analyze the data a lot of time and effort. Analyzing EMRs comes with its difficulties however. The EMR are conducted in an unstructured free-text format where much of the content revolves around attempts by the clinician to interpret whatever procedures the patient has gone through. Clinicians also differ in what terms they use to describe a patient's condition. They may use different synonyms for the same disease or different grammatical constructions entirely to describe the relationship between the different medical entities (Krzysztof & Moore, 2002).

EMR data can be divided into three different kinds: structured, semistructured and unstructured (Sun et al., 2017). Structured EMR data contains basic information such as birth data, medications, allergies and vital signs (height, weight, blood pressure etc). Semistructured EMR data usually take the form of a flow-chart and include name, value and time-stamps. Unstructured EMR data can be described as a one kind of narrative data, including documents such as medical notes, surgical records, discharge records and pathology reports. These unstructured texts include a lot of valuable information about patients but lack a common structural framework. Sun et al. (2017) further mention improper grammatical use, spelling errors, local dialects and semantic ambiguity as problems with the unstructured texts. Difficulties pertaining directly to the phenomena of keyword extraction can be further attributed to the relative scarcity of medical terms in the sentences that are being analyzed by the keyword extractor. Passages wherein a disease of interest are mentioned may be far and few between, as there is usually no need to repeatedly mention it. But if the patient currently is being treated for a specific disease that disease surely should be included as one of the keywords. Many keyword extractors that are based solely on certain statistical aspects of the document that is being analyzed will simply not realize the importance of this term. Therefore it is important to make sure that whatever keyword extraction method that is being applied to the clinical data is capable in understanding the context needed to find the most suitable keywords.

2.4 Evaluation metrics

2.4.1 Mean Average Precision

A commonly used evaluation metric for language technology is precision. Precision is the number of correct results divided by the total number of all returned results. When used for the purpose of evaluating automatic keyword extraction methods, calculating the precision provides the fraction of extracted keywords that are registered in the gold-standard. The problem with precision as an evaluation metric is that it treats keyword extraction as a classification problem. When evaluating keywords however, it is also of interest to consider the ranks of the candidate keywords. If a candidate keyword holds a high rank, it will appear earlier in the provided list of candidate keywords and should be considered differently from the candidate keywords with low ranks. The ranks of the candidates are not reflected in the evaluation result of the regular precision metric. However, an adaptation of the metric called the “Mean Average Precision“ (MAP) tries to address this issue by penalizing when irrelevant keywords appear with high rank in the candidate keyword list. MAP iterates through the candidate keyword list provided by the automatic keyword extractor and looks for the first relevant keyword it can find. When it finds a relevant keyword it creates a sub-list of the relevant keyword and all previous irrelevant keywords up until that point and subsequently calculates the precision score for that sub-list. MAP continues this process until it has finished iterating through the candidate keyword list and then calculates the mean average precision score for all the sub-lists. This provides us with a metric that also considers the ranking of keywords as part of the evaluation (Shrivastava, 2020).

2.4.2 Human evaluation and domain expertise

Human evaluations have long been important for applications of natural language processing models such as keyword extraction. After all, who is a better judge of language than the users themselves? Humans have the capability to understand fundamental aspects of texts which computers most often struggle to adequately solve, such as semantics. The use of humans for evaluations comes with its own limitations however, the obvious one being its time-consuming nature (Doddington, 2002). A somewhat less obvious limitation is the issue of domain expertise. When attempting to conduct studies within specific fields of inquiry a certain degree of domain knowledge is required (McCue, 2015). Domain expertise becomes even more important when trying to create applications with a target audience in mind. Asking a common man to evaluate a keyword extraction application that is meant to be used by medical professionals will yield different results than an evaluation performed by a team of domain experts. However, it is important to be aware of what sort of evaluation is obtained when using domain experts. What is of importance in the domain of dermatology when evaluating a keyword extractor will differ from that of cardiology.

3. Data

3.1 Electronic Medical Records

The data used in this thesis is a collection of EMRs that has been collected from two clinics at Linköping University Hospital; the cardiology clinic and the neurology clinic. The data has been collected through a period of five years and amount to a total of about 1 million EMRs collected from about 50 000 different patients. The cardiology clinic has contributed about two thirds of the total EMRs, with the rest contributed by the neurology clinic. The full distribution of EMRs can be seen in table 1. The EMRs vary in length, with some of them totalling less than 100 words while others may contain thousands.

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

Table 1: Number of tokens, EMRs and patients per clinic

3.2 Data preprocessing

Different keyword extraction models differ in their criteria for what type of preprocessing they require. Statistical methods, such as YAKE!, do not like noisy data. A substantial amount of the noise in the EMR dataset consisted of faulty sentence compositions caused by a lack of punctuation. For the YAKE! method, that puts a strong emphasis on how early in a sentence a word occurs, this sort of noise causes an obvious problem. Unsupervised learning methods that use embeddings on the other hand, such as KeyBERT, are less affected by noisy data. In some cases noise in the form of human errors like the misspelling of words might even be helpful in discovering the intrinsic properties of the text. However, to safeguard the result from inconsistencies that could be attributed to the noisy sentence compositions the keyword extraction methods were only applied on records where these errors did not occur.

3.3 Ethical considerations

Various precautions were taken to protect the integrity and privacy of the patients and doctors mentioned in the EMR data. Prior to gaining access to the data a confidentiality agreement was signed. The data could only be accessed through locally stationed computers or through a secure virtual connection to the local computers. The virtual connection required login credentials to the virtual environment and to the computer itself. The login credentials were handed out by local personnel at the hospital IT section. No files could be moved from the virtual environment to the host computer and the host computer had no access to the internet.

The content of the EMRs has not been fully anonymized and can therefore be used to identify patients or medical staff. Because of this, the data can not be shared with the public. Examples of keywords provided by the applications used in the thesis will nevertheless be discussed to some extent, given that the individual keywords do not include any identifiable information.

4. Method

4.1 Implementation of keyword extractors

Two different keyword extractors were applied on the EMR dataset; the statistical method YAKE! and the unsupervised learning method KeyBERT.

4.1.1 YAKE! implementation

A Python implementation of the YAKE! algorithm that had been distributed by the creators on github was used in the experiment (LIAAD, 2022). The parameters used were based largely on what was used in the original YAKE! thesis. The alternative to get rid of Swedish stopwords was used. Stopwords are common words who occur in all forms of texts but provide little to no meaning in and of themselves (Rajaraman, 2011). Some examples of Swedish stopwords are *och, att, så* and *av.* After some trial and error, the window size, the deduplication threshold and deduplication algorithm were all selected based on having shown the best performance in the thesis by the YAKE! creators, as the performance seemed to be negatively impacted when deviating from these values. In table 2 all the parameters that were given to the algorithm can be seen. The maximum length of the keywords (max ngram size) were picked based on observations made by the authors of the YAKE! thesis during their experiments.

Parameter	Value
language	sv
max ngram size	3
deduplication threshold	0.9
deduplication algorithm	seqm
window size	1
num of keywords	15

Table 2: YAKE! parameters

4.1.2 KeyBERT implementation

A Python implementation of KeyBERT distributed on github by the author of the original article was used in the experiment (Grootendorst, 2022). Two different BERT models were used for two different applications of the KeyBERT method. One of the applications used only the regular Swedish BERT model to create its embeddings while the other was also fine-tuned on the EMR data. In all other respects however both KeyBERT applications were identical.

The parameters that were used in the KeyBERT implementation were based on recommendations from the author of the algorithm (Grootendorst, 2021). The parameters can be seen in figure 3. Maximal Margin Relevance (MMR) was promoted by the author as a way to diversify the result of the application. MMR takes into consideration how similar a keyword is to already selected keywords. Depending on a diversity threshold value that range from 0 to 1, with 0 being not diverse at all and 1 being the most diverse, candidate keywords will be rejected if they are too similar to already selected keywords. A diversity value of 0.7 was hence chosen to help diversify the result of the extracted keywords.

The same stopwords, number of keywords (top n) and maximum length of the keywords (keyphrase ngram range) were used as in the corresponding values of the YAKE! implementation. This decision was made for two reasons: firstly, unlike the creators of YAKE! the author did not provide any guidance for what values might be best suited for the parameters; and secondly, to ensure that any differences in performance between the two implementations could only be attributed to intrinsic properties of each application.

Parameter	Value
use mmr	True
diversity	0.7
stopwords	YAKE! stopword list
keyphrase ngram range	1, 3
top n	15

Table 3: KeyBERT parameters

4.2 Human evaluation and creation of gold-standard data

Evaluation measures for keyword extraction usually rely on some form of gold-standard data that the extracted keywords can be compared to. As no such gold-standard data was available for the EMR dataset, part of the process involved trying to create it. This is a task that requires knowledge about the domain and the language used within it. To manage this, the gold-standard was created in collaboration with three radiologists via an human evaluation task.

For the human evaluation task 60 texts containing a patients full medical history were chosen from the EMR dataset. 30 of the histories came from the neurology data, 30 were picked from the cardiology data. The lengths of the patient histories varied from about 100 to 1000 tokens. From each patient history a set of candidate keywords were extracted using both the YAKE! application and the KeyBERT application. As these keywords are to behave as a form of baseline for further evaluations, the KeyBERT application using the regular Swedish BERT model without fine-tuning on the EMR data was chosen for the implementation. This also ensured that both the statistical approach and the unsupervised learning approach were represented in the human evaluation task. For each patient history, the applications extracted 15 candidate keywords each. As there is no real consensus on how many keywords should be extracted from any text of a certain length, 15 keywords were chosen to ensure that each history was provided with a healthy amount of keywords for the experts to evaluate. The extracted candidate keywords from each application were then merged to create a combined list where redundant keywords that were chosen by both applications were removed. Furthermore, candidate keywords that were part of larger compositions of keywords were also removed. This means that if for example both “broken left leg“ and “left leg“ were initially included as candidate keywords, only “broken left leg“ would make it to the human evaluation task. Both the merging and the removal of redundant keywords was done to make the tedious evaluation process a tad easier for the experts.

An excel file was created where each sheet corresponded to one patient history. After reading the patient history in full, the experts had to decide whether or not a corresponding keyword provided an important piece of information about the patient's health. Given their background in radiology, "important" information about the patient's health was interpreted by the experts as keywords with some relation to implants. They were then afforded three options; check a cell in the "YES" column for yes, in the "NO" column for no or in the "UNSURE" column if they could not decide. Finally, the experts were also prompted to provide their own suggestions of keywords from the patient history if they felt a word had been missed by the keyword extractors.

A gold standard of keywords for each patient history used in the human evaluation task was then created. Any keyword from the task that was rated positively by at least 2 of the 3 experts was included in the gold-standard dataset. To validate the ratings given by the experts an inter-rater reliability test was performed using Fleiss' kappa and Krippendorff's alpha. Fleiss' kappa measures the agreement between more than two raters where agreement due to chance is factored out (Fleiss, 1971). Krippendorff's alpha is similar to the kappa measurement but puts stronger emphasis on disagreement between the raters rather than agreement (Krippendorff, 2013).

4.3 Automatic evaluation

Using the gold-standard set of keywords created via the human evaluation task an automatic evaluation was performed. Each patient history in the evaluation dataset had 1-4 gold-standard keywords associated with them, with an average of 2 gold-standard keywords for each history. The patient histories were given to the keyword extraction applications that each extracted a ranked distribution of 15 candidate keywords. Each keyword in the ranked distribution were then compared to each gold-standard keyword to check for matches. A candidate keyword was considered a match if it was contained within the gold-standard keyword. This would for example match "left leg hurts" with "leg hurts" even though the two phrases are not a perfect match. These type of matches are called "PartOf" matches and were used in conjunction with "exact" matches. This type of approximate matching approach have been observed to be more similar to how human rating works (Zesch & Gurevych, 2009). A sublist of each candidate keyword up until a match was found would then be created. A precision score for this sublist was then calculated, before continuing to check for matches. When the list of candidate keywords had been either fully exhausted, meaning the application could not match all the gold-standard keywords, or each gold-standard keyword had been matched, an average precision score of the patient history was calculated. Finally, when all averages had been collected the MAP score was calculated.

5. Results

5.1 Results of manual evaluation

The individual ratings of each rater was compared using the Fleiss' kappa and Krippendorff's alpha based on 1152 keywords. The kappa score of 0.423 indicated a moderate agreement between the raters (Landis & Koch, 1977), the same indication was given by the alpha score (Krippendorff, 2011).

Keywords	Fleiss' kappa	Krippendorff's alpha
1152	0.423	0.424

Table 4: Inter-rater agreement

A full breakdown of each respective rater and their given ratings can be seen in table 5. As can be seen in the Y column, very few keywords were rated by the experts as giving important information about the patient's health. The amount of unsure responses, which can be seen in the U column, were also very low when compared to the N column. 44 keywords were rated U, while 39 were rated Y. Interestingly, the amount of given suggestions only amounted to 25 keywords, stemming from a total of 14 different histories, indicating that most of the patient histories did not include any important information at all in regards to their domain of expertise.

Rater	Y	N	U	Suggestions
1	9	1138	5	11
2	20	1111	21	6
3	10	1124	18	8

Table 5: All ratings per rater

The manual evaluation performed by the experts resulted in a total of 30 keywords which were rated positively or suggested by at least two of the three raters. Before including the keywords in the final set of gold-standard keywords, a minor cleaning process of the keywords was performed. A few instances of duplicates with very slight spelling differences were found in some of the subsets of keywords; for example "medtronic-pacemaker som impl" and "medtronic pacemaker impl". As to not penalize the extraction applications for essentially not extracting the same candidate keyword twice, the duplicate version with the most complex spelling were excluded from the final set of gold-standard keywords. In the case of the two examples given earlier, the latter would be included. The final total amount of keywords included in the gold-standard were 25, about 2 % of the total rated keywords. 14 of the 60 initial patient histories used in the evaluation had at least one gold-standard keyword associated with them and were thereby included in the set of patient histories used in the automatic evaluation. A full breakdown of the initial amounts of keywords and patient histories compared to what was then included in the gold-standard for the automatic evaluation can be seen in table 6.

Values	Initial amount	Gold-standard amount
Patient histories	60	14
Keywords	1152	25

Table 6: Results of manual evaluation

5.2 Results of automatic evaluation

The MAP score of each keyword extraction application can be seen in table 7. The observed map scores are very low, with both the KeyBERT applications only reaching a MAP score of about 4.5 %. The YAKE! application performed the best out of the three, although still only achieving a MAP score of 8 %. When only considering the patient histories for which the applications were able to successfully match at least one of the gold-standard keywords, the YAKE! application reached a MAP score of about 19 %. The KeyBERT application using the fine-tuned BERT model slightly outperformed the base model in both MAP scores, however, 90 % of the matches of the base model were exact matches while the fine-tuned version only reached an exact matching percentage of 37.5 %.

Application	MAP	Exact matches	PartOf matches	MAP excluding unsuccessful attempts
KeyBERT-base	0.043	9	1	0.08
KeyBERT-finetuned	0.048	3	5	0.10
YAKE!	0.081	4	7	0.1875

Table 7: Result of automatic evaluation

6. Discussion

6.1 Manual evaluation and creation of gold-standard keywords

The results of the manual evaluation were somewhat less extensive than first anticipated, as most of the patient histories that were used included no implant-related terms. However, the overall ratings of the experts were fairly similar. This indicates that an annotation task of this sort would be useful in trying to create gold-standard data. The importance of giving the rater the opportunity to essentially abstain from giving a definite answer was shown by the amount of unsure ratings. Rating keywords based on domain knowledge is a complex issue and allowing the raters to sometimes simply refrain from answering will solidify the integrity of the ratings given to the keywords that fall under the N or Y category. The lack of identified relevant keywords likely stem from the fact that the patient histories included in the evaluation in most cases had no association with the domain expertise of the three raters. This becomes apparent not only when considering how few keywords were rated as acceptable by the experts, but also when considering the small amount of suggestions given. As was touched upon in section 2.4.2, practitioners in different fields of medicine will interpret what is of importance in a patient history in different ways. The solution to this problem would most likely involve trying to first establish a gold-standard set of patient histories containing domain-specific terms of interest. A small step towards this has been achieved with the manual evaluation performed in this experiment, although a bigger set of patient histories would certainly be preferred in any future work.

6.2 Automatic evaluation, baseline and the limits of unsupervised learning approaches

None of the applications performed well in the automatic evaluation, with the YAKE! application slightly outperforming both the KeyBERT applications in terms of MAP scores. When only considering the patient histories for which the applications were successful in matching at least one gold-standard keyword, the MAP scores doubled for each application. The low scores likely reflect the main problem of using general models for domain-specific tasks, namely its inability to generalize to anything other than the main identity of the text. KeyBERT and YAKE! are both likely to do well in domain-specific tasks with a focus on terms directly related to the domain of the text, but will falter whenever the task deviates from that domain. The data used in these experiments, as previously mentioned in section 3.1, included patient histories from a cardiology and neurology clinic. For the implant-focused keyword extraction to be successful when using applications such as YAKE! and KeyBERT, the implants would likely have to be the main “ìssue“ covered in the patient history. Instead, as neither YAKE! or KeyBERT make use of any training data, the applications will tend to put their focus on attaining the most general distribution of keywords in regards to the texts content. In the end the YAKE! MAP score of about 19 % when only considering patient histories from which the application was able to successfully match at least one gold-standard keyword is likely the best contender for a simple first baseline of implant-focused keyword extraction.

6.3 Tendencies of the different keyword distributions

The distribution of keywords that were extracted differed somewhat between all three applications. KeyBERT-finetuned showed a bigger tendency towards including detailed information such as dosage when an opportunity to do so appeared. Misspelled words that were ignored by the other two applications were also very often extracted by KeyBERT-finetuned. This was likely a direct result of the finetuning on the EMR data, as neither KeyBERT-base or YAKE! showed the same tendency. In table 8, the top 3 extracted candidate keywords by each application for a singular patient history from the cardiology data can be seen. For KeyBERT-finetuned, two of three keywords is a combination of a misspelled word together with some form of medical information (I count the word "pat", short for patient, as a form of medical information). The misspelled words are hjäöortinfarkt (should be hjärtinfarkt, heart attack in english) and fäljande (should be följande, following in english). The word "pat", found in the second keyword of the KeyBERT-finetuned distribution, is an uncommon abbreviation of the word "patient" that neither KeyBERT-base nor YAKE! included in their distributions. In general, the two other applications would ignore misspelled words and uncommon abbreviations altogether. Even though the finetuning did not result in a performance that were significantly better than KeyBERT-base in this particular task, it seemed to show inclinations towards being better suited at recognizing distinct characteristics of the language used in EMR data. The YAKE! application would when possible lean towards including names of both patients and procedures in its distribution. This was likely a direct consequence of how the YAKE! model in general will give higher scores to words that include capitalized letters. As many of the gold-standard keywords were in fact related to different kinds of radiology procedures, this particular feature might have been a big contributing factor to why the application did better than the others at the task of finding keywords which the radiologists found to be indicative of giving important information about the patient's health. In general the three applications saw a decent amount of overlap in the type of words or phrases they extract. All three applications mainly provided distributions which contained medical information found in the text, albeit with slight differences. KeyBERT-base provided a more general distribution of keywords that more often than not included some sort of medical information. KeyBERT-finetuned on the other hand put stronger emphasis on singular medical terms with a more protruding nature. Finally, the YAKE! application leaned towards providing a distribution that contained names of both people and procedures.

Application	Keywords
KeyBERT-base	ökat symptombild, faxat, ioleostomi diskuteras
KeyBERT-finetuned	hjäörtinfarkt hand behandlas, pat tagit fäljande, beskriver ökande anginösa
YAKE!	ileostomi pga tarmischemi, träffar patienten, avdelning efter colektomi

Table 8: Top 5 ranked keywords of each application for a singular patient history

6.4 Possible improvements of the applications

The results of the YAKE! application is unlikely to be improved upon as the statistical underlining of the model will remain the same regardless of any finetuning done to its parameters. It might be possible to impose certain restrictions on what kind of keywords are to be allowed in its final distribution by, for example, incorporating more linguistic knowledge or matching them with a list of implant-focused terms. This would defeat the purpose of using an unsupervised approach, and in this case you might as well just make use of a supervised approach instead. The results of the KeyBERT applications however could likely be improved without the introduction of explicit supervision. As described in section 2.2.2, KeyBERT utilizes BERT-embeddings to guide its final distribution of keywords towards ones that are similar to the document as a whole. In addition to this document embedding, a set of seeded terms could be given to the model. The model will then be instructed to not only consider the document embedding when selecting candidate keywords, but also the embeddings provided by the set of seeded terms. If a set of implant-related terms were available to use for this purpose it would likely change the results for the better.

7. Conclusions and future work

Clinical data remains a rather unexplored area of automated text, implant-focused keyword extraction even less so.

The first research question of this thesis inquired about potential limitations of unsupervised keyword extraction approaches when posed with the domain-specific task of implant-focused keyword extraction from EMRs. The results of KeyBERT and YAKE! applications indicate that the main limitation of these approaches for this type of task is also what is regarded as their main strength; their ability to capture the general content of a text. Unless the implants are closely related to the main content of the text, the applications will not include them in their distribution of candidate keywords. As the data that was used for the experiments were by and large “about“ something different from the implant-related terms found in the text, basic implementations of unsupervised approaches will fail in recognizing them as important. Slight adjustments to the KeyBERT application in the use of seeded terms will probably improve upon its ability to extract implant-related keywords, whilst the YAKE! application is unlikely to be improved at all.

The second research question asked whether a general language transformer model, like the Swedish BERT, can be applied to domain-specific keyword extraction. The results of the experiments with KeyBERT, albeit stemming from a small sample size, indicate that a general model such as Swedish BERT perform very similarly to a BERT model that has been finetuned on domain-specific data. Neither the finetuned model nor the general model performed well when tasked to extract implant-focused terms from patient histories in which the implants only held a loose association to the main content of the text. The final distributions of keywords for both models however give some indications of being better suited in dealing with simpler domain-specific tasks such as extracting heart-related terms from cardiology data.

The third and final research question was aimed at exploring the possibilities of using unsupervised approaches as a crutch in creating a domain-specific gold-standard of keywords. KeyBERT and YAKE! applications were utilized to select keywords for an annotation task that were to be performed by three radiologists. The results of the annotation task indicated that the unsupervised applications were able to include about two thirds of the implant-related terms detected in the patient histories. The rest had to be included as suggestions by the radiologists themselves. The ratings given by the three radiologists showed a moderate agreement

This thesis provided the basic foundational groundwork for the complex issue of clinical keyword extraction; the limitations of unsupervised approaches, the applicability of general language transformer models such as Swedish BERT to domain-specific tasks and finally the problems associated with creating gold-standard data. Further research is needed however, perhaps involving the possible modifications of KeyBERT discussed in chapter 6.4. Additionally, research that helps establish a larger set of gold-standard for patient histories containing implant-related keywords would strengthen the foundational work provided here. This will in turn open up possible avenues to areas involving supervised keyword approaches, enabling an enhanced understanding of how to handle clinical data.

Bibliography

Beliga, S., Mestrovic, A., & Martincic-Ipsic, S. (2014). Toward Selectivity Based Keyword Extraction for Croatian News. *arXiv preprint* arXiv:1407.4723v1.

Beliga, S., Meštrović, A., & Martinčić-Ipšić, S. (2015). An overview of graph-based keyword extraction methods and approaches. *Journal of information and organizational sciences, 39*(1), 1-20.

Bracewell, D. B., Ren, F., & Kuriowa, S. (2005). Multilingual single document keyword extraction for information retrieval. *International Conference on Natural Language Processing and Knowledge Engineering*, 517-522.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences, 509*, 257-289.

Cios, K. J., & Moore, G. W. (2002). Uniqueness of medical data mining. *Artificial intelligence in medicine, 26*(1-2), 1-24.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint* arXiv:1810.04805.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Proceedings of the second international conference on Human Language Technology Research*, 138-145.

Firoozeh, N., Nazarenko, A., Alizon, F., & Daille, B. (2020). Keyword extraction: Issues and methods. *Natural Language Engineering, 26*(3), 259-291. doi:10.1017/S1351324919000457

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378-382.

Grootendorst, M. (28 October 2020). Keyword Extraction with BERT. *Maartengrootendorst.*

https://www.maartengrootendorst.com/blog/keybert/

Grootendorst, M. (2022). KeyBERT. https://github.com/MaartenGr/KeyBERT

Han, J., & Kamber, M. (2006). Classification and prediction. *Data mining: Concepts and techniques, 2006*, 347-50.

Jerdhaf, O. (2021). *Discovering Implant Terms in Medical Records.*

Krippendorff, K. (2011). Computing Krippendorff's Alpha-Reliability.

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology.* (3rd edition). Sage.

Labaratory of Artificial Intelligence and Decision Support. (2022). Yet Another Keyword Extractor (Yake). https://github.com/LIAAD/yake

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics (33)*, 159-174.

Leitner, F. (2021). *Sentence segmentation and word tokenization.* https://pypi.org/project/segtok/

Malmsten, M., Börjeson, L., & Haffenden, C. (2020). Playing with Words at the National Library of Sweden–Making a Swedish BERT. *arXiv preprint* arXiv:2007.01658.

McCue, C. (2015). *Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis.* (2nd edition). Butterworth-Heinemann.

Perlis, R. H., Iosifescu, D. V., Castro, V. M., Murphy, N. S., Gainer, V. S., Minnier, J., Cai, T., Goryachev, S., Zeng, Q., Gallagher, P. J., Fava, M., Weilburg, J. B., Churchill, S. E., Kohane, I. S., & Smoller, W. J. (2012). Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychol Med, 42*(1). doi:10.1017/S0033291711000997.

Rajaraman, A., & Ullman, J. D. (2011). *Mining of massive datasets.* Cambridge University Press.

Roden, D. M., Pulley, J. M., Basford, M. A., Bernard, G. R., Clayton, E. W., Balser, J. R. & Masys, D.R. (2008). Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. *Pharmacoethnicity, 84*(3), 362-369. https://doi.org/10.1038/clpt.2008.89.

Shrivastava, I. (14 July 2020). Exploring Different Keyword Extractors - Evaluation Metrics and Strategies. *GumGum Tech.* https://medium.com/gumgum-tech/exploring-different-keyword-extractors-evaluation-metrics-and-strategies-ef874d336773

Siddiqi, S., & Sharan, A. (2015). Keyword and keyphrase extraction techniques: a literature review. *International Journal of Computer Applications, 109*(2).

Sun, W., Cai, Z., Liu, F., Fang, S., & Wang, G. (2017). A survey of data mining technology on electronic medical records. *IEEE 19th international conference on e-health networking, applications and services (Healthcom)*, 1-6.

Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., & Wang, G. (2018). Data Processing and Text Mining Technologies on Electronic Medical Records: A Review. *Journal of Healthcare Engineering.* https://doi.org/10.1155/2018/4302425

Tang, M., Gandhi, P., Kabir, A. M., Zou, C., Blakey, J., & Luo, X. (2019). Progress Notes Classification and Keyword Extraction using Attention based Deep Learning Models with BERT. *arXiv preprint* arXiv:1910:05786v2.

Turney, P. D. (2000). Learning algorithms for keyphrase extraction. *Information retrieval, 2*(4), 303-336.

Zesch, T., & Gurevych, I. (2009). Approximate matching for evaluating keyphrase extraction. *Proceedings of the International Conference RANLP-2009.* 484-489.