Knowledge Extraction and Prediction from Unstructured Medical Documents

Ioana Barbantan, Rodica Potolea


Abstract. Concept extraction from unstructured documents is a sensitive step in the knowledge extraction process. The transformation of the unstructured documents into structured data and further into structured information that can be exploited by predictive modules implies a sequence of steps related to data preparation, concept extraction and data organization, filtering and selection. Along with the adoption of the Electronic Health Records a step forward has been taken in the healthcare domain allowing for faster and easier knowledge discovery and utilization of data. While the ultimate goal for structuring EHRs is knowledge extraction for building assistive medical diagnosis systems, the correct identification of terms in documents is essential. The current paper covers an original complete solution for automatically structuring medical documents and extracting relevant medical concepts via the PreNex [1] and MedCIM [2] strategies while our vision for the Knowledge Extraction and Prediction solutions is being argued and is under development.

Keywords: hybrid algorithm, prediction, recommendation, Electronic Health Records, Data Mining, Text Mining

1 Introduction

Making assumptions about healthcare conditions is a common concern for every individual. Several desktop and mobile applications help us identify the conditions that bother us. The input data fed into the applications is represented by the symptoms we report, while the output lists suggestions concerning the conditions that might affect us. The major drawback of these applications is that they may not use authorized content and comprehensive knowledge sources. However, a source of trustworthy data is included in the Electronic Health Records, which we will refer to in the following as EHRs. They enable identifying the relationship between medical concepts, predicting epidemics or detecting cases of rare diseases [3]. Domain oriented ontologies like SNOMED-CT [4] or UMLS [5] encodes the information captured in the EHRs that enable making predictions driven by assistive decision support systems.

Three parameters carry significant importance when dealing with healthcare management systems: <patient suffering, medical costs, and time>. Making the
tradeoff between these parameters while providing an accurate diagnosis or deciding if no further investigations are needed, are decisions that the medical doctors make every day while being constrained by time. The existence of a knowledge base of former patients that were previously investigated and diagnosed, benefits along all dimensions: health condition, costs, diagnosis, and hospitalization time. Therefore, the ultimate goal is obtaining recommendations provided by an assistive decision support system employing such a knowledge base. The consequences are the decrease of patient’s suffering and a reduced number of medical investigations, qualified both as costs and time interval between the patient's hospitalization and start of treatment, thus initiating the healing process.

The rest of the paper is organized as follows. In chapter II we present similar solutions proposed for handling the information extracted from the EHRs. In chapter III the conceptual approach for making predictions from discharge notes is introduced, while in chapter IV an instantiation of such a solution is presented. The last chapter includes the conclusion of our work and future enhancements for our proposed approach.

2 Related Work

Building an assistive decision support system using predictive analysis requires structured data and a solid knowledge base. However, practical issues concerning the acquisition of the structured data arise. While the input data is represented by unstructured documents, namely, EHRs, several text preparation steps are required to transform the unstructured text into structured data. To deal with the data transformation we proposed the MedCIM [2] strategy which consists of a sequence of text transformation steps followed by token filtering, similarity measurements, semantic interpretation, and finally category identification.

The increasing trend of EHRs adoption is reported by the authors in [6] by analyzing the trends between 2007 and 2012. While the adoption rate was 34.8% in 2007, it increased 37%, reaching 71.8% in 2012, thus showing the necessity of solutions to analyze, exploit and extract knowledge from the data.

To tackle data prediction and correlation, several supervised and unsupervised [7] methods were approached thus identifying the relationship between the concepts in a medical document. The relationships can be established between a disease and the medication administered or can help in identifying cases of adverse effects. The same research group proposed a supervised approach for identifying relationships in [8]. The correlation of the concepts is computed using additional external resources such as Wikipedia.

The study in [9] proved that the information encoded in the EHRs can help in predicting the evolution of the health status of ward patients. Based on the patients' vital signs and laboratory results, the authors have shown that for each patient a prediction can be made: either the patient suffered from cardiac arrest or would be transferred to the ICU.

The authors in [10] proposed two statistical measures, the chi-square and the proportion confidence interval (PCI) method, to measure the dependence of pairs of
diseases and findings. Their research showed that co-occurrence was more relevant than collocations when trying to establish the disease-finding relation.

In a predictive model, as the matrices used for making the predictions are sparse and have large numbers of features, and are rather unbalanced, specialized Machine Learning techniques are required to manage them. The authors in [11] developed the ProICET strategy to tackle the medical diagnosis and prognosis from such matrices.

3 Conceptual Approach for Making Predictions from Medical Documents (EHRs)

From the data perspective, the EHRs are sources of unstructured information while from the medical perspective they are valuable sources of information conveying medical data concerning patients’ health status, the conditions they suffer from, medication they follow or the procedures underwent. EHRs focus on all medical aspects of a patient’s health and help identifying correlations between the current and previous investigations and conditions [12].

However, during a medical examination the patients may be required to fill in an initial report about their health status. In [13], studies dealing with the agreement level between what the patients report and the information they further share with the medical doctors and the outcome of the investigations, proved that a self-assessment is not accurate and the existing medical recommendation systems can only be used as a guideline but not as a source of truth.

Making predictions about a patient’s health status is a task that many researchers tackle employing Machine Learning techniques or pattern matching algorithms [7], [8], [10]. The challenges faced when developing an assistive decision support system are manifold and can be organized into the following categories: data preprocessing, extraction, structuring, and prediction, as depicted in Fig. 1. The predictive models can help identifying patients with high risks, thus providing ahead of time the necessary investigations and treatment.

![Fig. 1. General flow for making predictions from unstructured documents.](image)

3.1 Natural Language Processing

The general flow of a concept mapping approach consists in the following steps. The solution takes as input any type of unstructured document and performs several text processing and linguistic analysis steps. The two steps are responsible with
converting the unstructured text into structured data enhanced with semantic information. Several common tasks are considered: part of speech (POS) identification, stop words removal and concept normalization via a lemmatization step.

3.2 Concept Extraction

Extracting the relevant medical concepts from EHRs implies discriminating between the usual terms and the relevant concepts to the medical domain. Establishing the association between the relevant concepts, the medical domain, and the corresponding semantic categories is supported by external sources. The concept mapping task deals with identifying the textual references of the concepts of interest which are annotated with specific entries of lexicons and terminologies. The output generated at this step can be represented by a sparse $A_{m\times n}$ matrix. In case the goal is suggesting diseases based on the findings and the symptoms reported by the patient, the matrix is represented as follows. The columns are symptoms while the lines correspond to the structured representation of an EHR. The value of the element at position $(i,j)$ - $a_{i,j}$ indicates the status of the symptom at position $j$ when the patient is examined based on the EHR at position $i$. The possible values for the element are presented below, in Equation 1.

$$a_{i,j} = \begin{cases} 
0, & \text{if the concept does not exist in the EHR or is negated;} \\
1, & \text{if the concept exists in the EHR;} \\
x, & \text{if the concept exists and can be quantified.} 
\end{cases}$$  \hspace{1cm} (1)

When assessing a patient’s health status, the medical doctor considers several factors such as the medical history of the patient, the symptoms they report, the medication they are under, the procedures underwent or follow-up appointments. In most medical documents the information concerning these factors is unstructured. Procedures and medication related information can not only be found in the medical history section, but also in the corresponding sections. Thus, a second challenge faced when developing an assistive decision support system is obtaining structured information such that it becomes easily accessible. The input received from the previous step - the identified medical concepts and the semantic classes they are associated with – leads to a possible restructure of the documents such that the all sentences and phrases belonging to a semantic category are grouped.

3.3 Making Predictions

The target module concerns inferring knowledge from the previously structured data. When the documents are structured and the information is normalized, a dedicated system makes predictions based on a knowledge base consisting of patients with similar conditions that were treated previously. The predictions are medication recommendations, the disease they may suffer from, investigations, or the evolution
of the patient’s health status. The final goal is to infer such decisions for a new patient based on patients with similar conditions.

4 Instantiating the Conceptual Approach for Making Predictions from Discharge Notes (EHRs)

The final goal of our work is knowledge extraction leading to an assistive decision support system. To achieve our goal we developed strategies to cover the main modules; an NLP tool for data preparation, the MedCIM strategy [2] for concept identification, depicted in Fig.2, and we our current work focuses on the prediction module.

![MedCIM flow](image)

**Fig. 2.** MedCIM flow.

4.1 Natural Language Processing

Similar to a typical Text Mining problem, the commonly used preprocessing steps are employed; a tokenizer to identify the tokens in the document, a lemmatizer to normalize the word tokens, a POS tagger to be further used in filtering the tokens and an n-gram selector. While for the tokenizer and n-gram selector we create our own implementation, the POS tagger employed is the Stanford POS Tagger [14] and for identifying the lemma of the terms, we employ the WordNet lemmatizer [15].

According to the analysis we performed in [1] on the MTSamples dataset [16], negation in EHRs plays a significant role as 2% of the total number of words used in filling in the EHR are negated, and moreover, the distribution of syntactic and
morphologic negation is comparable. To tackle the syntactic negation identification task in medical documents, in particular, the NegEx [17] system was developed. To the best of our knowledge, morphologic negation is not treated in the existing systems dealing with negation identification and exploitation, and that is why part of or research was conducted in this area.

Our previous work [1], [18] deals with treating morphologic negation - negation expressed using negation prefixes and proposed the BOW-NPI [18] and the PreNex [1] strategies. The two strategies differ in the type of external data sources included in the analysis; the first uses a corpus-dependent terminology list while the second uses the English dictionary. The list of negation prefixes for the medical domain is rather limited. The negation prefixes used in our analysis are: a, an meaning not or without, anti (opposing, against), dis (negation, removal, expulsion), in, il, im, ir (not, without), non (absence, negation), un (not, reversal, cancellation, deprived of).

The two strategies consist of several linguistic rules that cover the cases identified for morphologic negation and presented in the following. To determine whether a word is prefixed with negation prefix, we need to identify whether the root of the word is actually a word. The BOW-NPI [18] strategy requires a preliminary terminology building step. A set of documents are used to create a bag of words representation that will be further used as knowledge source. Our second approach, PreNex, follows a similar strategy as BOW-NPI but extends the terminology source to the entire list of words in the English language by querying the WordNet dictionary. As the NegEx algorithm is specialized in identifying explicit negation occurrences, its performance in terms of recall is limited, when including the negation expressed with prefixes in the negation identification phase. As shown in Table 1, the first line, the recall for NegEx is only 53.14%. The improvement that is obtained by including the identification of our proposed solution for prefixed negation identification is of 39.98%. Although there is a slight decrease in the value of the precision in the overall negation identification solution, the increment obtained for recall makes the solution adequate for the problem we solved.

<table>
<thead>
<tr>
<th>Method employed</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>NegEx</td>
<td>98.76%</td>
<td>53.14%</td>
<td>69.1%</td>
</tr>
<tr>
<td>PreNex</td>
<td>96.61%</td>
<td>93.03%</td>
<td>94.78%</td>
</tr>
<tr>
<td>Improvement NegEx → PreNex</td>
<td>-2.15%</td>
<td>39.89%</td>
<td>25.68%</td>
</tr>
</tbody>
</table>

In summary, Table 2 displays the comparable performance of the two proposed strategies, and shows that extending the knowledge source from a closed domain specific list, to the entire list of words existing in a language, improves the negation recognition rate (an F-measure increment of 3.56% is obtained).

<table>
<thead>
<tr>
<th>Method employed</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOW-NPI</td>
<td>95.79%</td>
<td>87.63%</td>
<td>91.53%</td>
</tr>
<tr>
<td>PreNex + Syntactic Negation</td>
<td>95.96%</td>
<td>94.23%</td>
<td>95.09%</td>
</tr>
</tbody>
</table>
To prove the strength of the PreNex approach, we proposed the RoPreNex cross-language approach [19]. Our specific goal was to instantiate a cross language methodology that identifies morphologic negation in both the source and target languages using the linguistic resources in each corresponding language. The source language is English, for which we have proposed and evaluated the PreNex strategy and the target language is Romanian, which represents the subject of the approach. The results reported for identification are promising and to our best knowledge, there are no similar approaches for identifying morphologic negation for the Romanian language. The results obtained by the proposed methodology report precision of 77.78% and recall of 80.77%, which are only 14.85% in case of precision and 12.83% in case of recall below the corresponding English results.

4.2 Concept Identification in Unstructured Medical Documents

For the task of concept identification in unstructured medical documents we propose the MedCIM concept mapping strategy that consists of two separate flows. The first is related to processing the raw document and identifying the medical related candidates, while the second flow is related to mapping the candidates to the medical terminology, and therefore identifying the actual medical concepts as shown in Fig. 2.

The document flow follows the steps that are applied on the raw document (detailed in section 4.1) up to the point of identifying the possible medical candidate terms. To identify whether a concept is a medical concept candidate, we validate its relationship with the semantic medical classes defined in WordNet. The possible medical candidates we propose are the terms and n-grams which are included in the medical WordNet classes and the terms that do not have a definition in WordNet. The WordNet filtering step is followed by the mapping process, where for each medical concept candidate we search for the corresponding instance in the SNOMED-CT ontology, extract the values of its properties that will be further preprocessed. We extract from the ontology the instances included in the semantic categories defined in UMLS [20], further grouped into: symptoms (Sign or Symptom T184, Cell or Molecular Dysfunction T049), diseases (Disease or Syndrome T047, Pathologic function - T046, Mental or Behavioural Dysfunction T048, Congenital Abnormality T019, Acquired Abnormality T120, Anatomical Abnormality - T190, Neoplastic Process T191, Experimental Model of Disease T050), procedures (Health Care Activity T058, Laboratory Procedure T059, Diagnostic Procedure T060, Therapeutic or Preventive Procedure T061), and medications (Pharmaceutical Substance T121, Organic Chemical T109, Clinical Drug - T200, Antibiotic T195).

The approach is quantity (recall) driven, as it is more important in the medical diagnosis problems to identify all the possible medical conditions and symptoms of a patient, even though some become unnecessary, rather than missing a relevant finding that could make the difference between a patient that is sent home ill and a patient that is having additional tests when the preliminary investigations do not lead to an obvious diagnosis. Moreover, it is more feasible and efficient for a medical doctor to eliminate the false positive concepts than to try to identify the overlooked medical concepts. The outcome of the initial evaluation of the MedCin solution showed that some categories of medical concepts were missed, namely, the medical concepts
expressed using abbreviations and morphologic negation. To handle these special cases we included an additional external resource, a concept list, containing the abbreviations with their meaning and the PreNex identification strategy. Successively handling these special cases, the outcome of the MedCIM strategy showed the identification of medical concepts with a recall rate of 74.83% and a precision of 73.7% (F-measure of 74.26%), as shown in Table 3, last line.

Table 3. Evaluation of category identification on the 2010 i2b2/VA Challenge train dataset[21]

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Initial Evaluation</td>
<td>72.21%</td>
<td>67.1%</td>
<td>69.56%</td>
</tr>
<tr>
<td>2. 1 + Handling Abbreviations</td>
<td>73.43%</td>
<td>72.65%</td>
<td>73.04%</td>
</tr>
<tr>
<td>3. 2 + Handling Morphologic Negation</td>
<td>73.7%</td>
<td>74.83%</td>
<td>74.26%</td>
</tr>
</tbody>
</table>

A thorough analysis on the errors produced by the proposed approach shows that improvements can be made to reduce the error rate. One source of errors is the incomplete concept coverage in SNOMED-CT which does not include disease related concepts such as orthostasis, voluntary guarding or the trade names of the drugs (e.g. Lupron, Protonix or Levaquin). While a limited number of trade names were identified using the WordNet dictionary, for the others a solution can be integrating additional external resources which includes the commercial names for medication. Other false negatives have been produced by medical abbreviations such as creat, while in the symptoms category, terms such as thickened and redness, are overlooked as are used frequently in common language, but become context dependent medical concepts. Another aspect which we are currently considering is the ability to distinguish whether a body part is included in a disease description or it is referring to the actual anatomical structure.

4.3 Knowledge Extraction and Prediction

The final goal of our work is knowledge extraction leading to an assistive decision support system designed as a personal medical assistant application. The main steps to consider for reaching the desired outcomes are EHR structuring, knowledge extraction and data prediction. As the EHRs are unstructured documents, they carry information, but if exploited correctly they become valuable sources of knowledge. By imposing a structure, the EHRs are transformed into structured data that can be further used to infer new data - knowledge.

Most of the research presented in this paper deal with the data preparation steps, thus obtaining the structured data from the EHRs. After identifying the relevant concepts and associating them with semantic classes, we further need to enhance the knowledge base by quantifying the extracted concepts. The disease related concepts have binary quantifiers (present or absent/negated) while the symptoms are quantified as both binary features or carry a distinctive value (like in the case of the temperature that can be expressed in a range of values). Other such quantifiable concepts are medication and laboratory tests that are evaluated as numerical values. Every EHR is linearized and represented as a feature vector. Each element in the array -feature
vector- carries the value of the measured concept. The collection of all the feature vectors that are our knowledge base is represented as a sparse matrix that can be fed to a learning algorithm. The knowledge base is organized into use-case-oriented domains, namely 1) disease prediction for a patient based on the previously investigated patients and 2) suggestion of the limited number of investigations to be performed to remove uncertainty, increase prediction rate, reduce the number of invasive procedures a patient undergoes, having implications in reducing the medical costs, also. A clustering algorithm is employed to automatically identify the clusters of patients with similar conditions. When for a new patient a prediction is required, his health status is fed to the previously generated model and based on the similarity of his condition and the content of the knowledge base suggestions can be made.

5 Conclusions

Our strategy aims to be a step towards a medical assistive decision support system: starting from raw medical data, it infers the appropriate suggestion to each specific task (further investigations, diagnosis or medication). The medical documents which are usually stored in unstructured format can be structured using a terminology mapping technique. The required preprocessing steps proved to have a significant role in normalizing both the input text (unstructured data) and the terminology sources (structured data). The filtering step which discriminates between medical and non-medical concepts via the WordNet dictionary proves to be an efficient method for filtering the non-medical concepts. In the selection of the terminology sources (WordNet and SNOMED-CT) their ability to cover the biomedical domain and also to obtain accurate information was considered.

The current status covers complete solutions for automatically structuring medical documents and extracting relevant medical concepts via the PreNex and MedCIM strategies while the Knowledge Extraction and Prediction tools are under development.

Acknowledgments

The authors would like to acknowledge the contribution of the COST Action IC1303 - AAPELE.
This work has been partially supported by Brained City - Information Technology based Innovative Development of Cluj-Napoca Fully Integrated Urban Ecosystem.

References

3. Fung, K.W., Richesson, R., Bodenreider, O.: Coverage of Rare Disease Names in Standard Terminologies and Implications for Patients, Providers, and Research. AMIA Annu Symp Proc (2014)