THE TOOLS AND RESOURCES FOR CLINICAL TEXT PROCESSING

Ulfeta Marovac*, Aldina Avdić

Department of Technical Science, State University of Novi Pazar, Novi Pazar, Serbia

Correspondence:
Ulfeta Marovac
umarovac@np.ac.rs

Abstract:
The expansion of electronic health has produced a large amount of information in medical information systems in a structured and unstructured format. The processing of unstructured data in the form of text is performed using natural language processing techniques. Natural language processing requires specific resources for processing text depending on the domain of the text as well as the language in which it is written. This paper aims to present the available tools, lexical resources, and corpora used in the analysis of clinical texts.

Keywords:
natural language processing, clinical texts, lexical resources.

INTRODUCTION

Most countries today use e-health and keep large amounts of data about patients and their illnesses that are usually used to manage individual cases as well as administration. Data stored in medical information systems can be structured and unstructured. Unstructured data contain free text such as anamnesis, radiological reports, patient care notes, and other similar clinical texts. This data carries information that is important not only for resolving an individual case of the disease but also for extracting general information. Health care is one of the top priorities of every state. We are witnessing that diseases have no borders and language barriers and that the fast and efficient availability of data can save people's lives. Most developed countries are seriously processing information from clinical texts to improve health, but most of the research is for the English-speaking world. Creating standards for storing medical data in a multilingual system could contribute to the faster development of medicine. If there was a single platform through which COVID-19 patients would be monitored, it would be possible to get information...
about places of potential hotspots of infection [1], clinical characteristics and prognostic factors using natural language processing methods [2]. The ability to process reports of patients treated in countries where English is not an official language allows for global aggregation of data, which is extremely important especially for rare diseases [3]. This paper provides an overview of available resources for processing clinical texts in different natural languages. The paper is organized into six sections. The second section presents a description of the basic concepts on which the work is based. The third section shows the processing of the clinical text. An overview of the corpora used for different languages is given in the fourth section. The fifth section contains tools for medical classification and annotation of clinical texts. The last section contains the conclusion and directions for further research.

2. THE BASIC CONCEPTS

Introduce the basic concepts of clinical text processing that are the main points of consideration in this paper: clinical text, electronic medical reports, natural text processing, clinical text processing, resources for clinical text processing, and medical classification.

Clinical texts are written by doctors, medical staff, and other health care providers. They are used to document the patient’s condition and the health services provided. They describe patients, their pathologies, their personal, social, and medical history. Clinical texts differ from scientific texts and are not prepared for publication. They do not use complete sentences, they use medically accepted expressions and abbreviations that do not belong to the natural language in which they were written.

Electronic health records (EHR) carry a lot of important information such as the patient’s condition on admission to the hospital, the course of his recovery, then the state of health at discharge. This information is still most easily expressed in natural language, which makes information extraction more difficult [4]. Specific medical terminology is defined by different standards and classification systems. Classification and descriptions of diseases, treatments, and drugs control the vocabulary used in medical reports and administration reduce ambiguity and increase the degree of clarity.

Natural language processing (NLP) is a field of linguistics, computing, and artificial intelligence that explores ways in which computers can understand and use text or speech in natural language and apply them to some useful activities [5].

Clinical text mining is the extraction of information from clinical texts [6].

Resources for processing clinical text are all data sets that help in research, and they can be: sets of diagnoses, symptoms, drugs as well as clinical corpora.

Medical classification and terminology are classification systems and terms used in reports, administration, classification and description of diseases, treatments, and medications such as ICD coding diagnosis, SNOMED CT, MeSH, UMLS, ATC, and others [6].

3. THE CLINICAL TEXT PROCESSING

Clinical texts represent the basic form of communication between healthcare professionals. Using natural language processing methods, it is possible to extract information from these texts that are hidden in free text and which are not easily usable for computer analyzes. Most of the authors are engaged in the analysis of data from clinical texts written in English due to their public availability as well as the public availability of clinical text processing tools for English. Two approaches are used to process natural language in clinical texts: a rule-based approach; and machine learning algorithms. The first approach requires the existence of specialized clinical dictionaries that support complex clinical logic such as the MTERMS tool [7].

The second approach based on machine learning requires a set of manually annotated clinical data. An overview of the use of machine learning over clinical texts until 2020 is given in [8], where the results of 110 research available on PubMed from the period from 2015 to 2018, which concern the machine processing of clinical texts in English, are presented. Examining the properties of the data used, it was concluded that most of the research used hundreds or thousands of documents.

There is a small number with a very small data set of less than 50 and a very large of 10,000 documents (ten papers). Many of the data, although available, remained unused. The main reason for unused data is that they are not marked. If the data annotation is done manually then it is a hard job and prone to errors. Active learning algorithms enable the processing of documents even with a smaller number of manually annotated data, whereby new annotations use more algorithms and compare their results. Existing structured data are often used for annotation, so the textual part of the medical report can be labeled using the diagnosis code [9].
Semi-automated annotation can also be used. Often the data being processed comes from one institution, so the relevance of that data is questionable. Very often the results published on one data set did not give the same results on another set [10]. The clinical application of such data processing is wide from diagnostics, prognosis, protection, risk prediction, improvement of service provision, management, etc.

4. AVAILABLE CLINICAL CORPUS

It is very difficult to reach the clinical corpus due to the sensitivity of the data it contains. Each clinical corpus must have ethical approval for use. The data must pass the deidentification process to preserve patient privacy by taking into account names and identification numbers, telephone numbers, and addresses. There are several corpora available for both English and other languages.

For English, more data sets are annotated and consist of discharge lists, medical histories, nursing reports, radiological reports, sentences from the medical domain, and other medical reports. Some of them are available:

- Informatics for Integrating Biology & the Bedside (i2b2) [11]
- Computational Medicine Center (CMC) corpus [12]
- ShARe/CLEF eHealth [13]
- Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) [14]
- BioScope Corpus [15].

There are clinical corpora in non-English languages, but they are smaller and cover fewer different medical contents. The number of publications in the field of natural text processing in different languages in PubMed is shown in the graph (Figure 1).

Non-English corpora are commonly from the researcher’s institutions and require special permits and contacts for their further use. These corpora contain labels for: diagnoses, symptoms, medications, and therapies. Some of the corpora of clinical texts in non-English languages used in the scientific papers are presented in Table 1.

<table>
<thead>
<tr>
<th>Language</th>
<th>Description</th>
<th>No. of texts</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>discharge reports</td>
<td>142 154</td>
<td>[16]</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>clinical texts</td>
<td>100 000 000</td>
<td>[17]</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>outpatient records (diabetic)</td>
<td>500 000</td>
<td>[18]</td>
</tr>
<tr>
<td>Serbian</td>
<td>medical records (B05)</td>
<td>5000</td>
<td>[19]</td>
</tr>
<tr>
<td>Serbian</td>
<td>medical reports</td>
<td>4212</td>
<td>[20]</td>
</tr>
<tr>
<td>Serbian</td>
<td>medical documents</td>
<td>200</td>
<td>[21]</td>
</tr>
<tr>
<td>Swedish</td>
<td>Stockholm EPR Corpus</td>
<td>2 000 000</td>
<td>[22]</td>
</tr>
<tr>
<td>Danish</td>
<td>clinical narrative text</td>
<td>61000</td>
<td>[23]</td>
</tr>
<tr>
<td>Dutch</td>
<td>EMC Dutch clinical corpus</td>
<td>-</td>
<td>[24]</td>
</tr>
<tr>
<td>Finnish</td>
<td>intensive care nursing narratives</td>
<td>2800</td>
<td>[25]</td>
</tr>
<tr>
<td>French</td>
<td>clinical texts</td>
<td>170 000</td>
<td>[26]</td>
</tr>
<tr>
<td>Italian</td>
<td>clinical texts</td>
<td>23 695</td>
<td>[27]</td>
</tr>
<tr>
<td>Italian</td>
<td>clinical texts</td>
<td>100</td>
<td>[28]</td>
</tr>
<tr>
<td>German</td>
<td>clinical texts</td>
<td>18 000</td>
<td>[29]</td>
</tr>
<tr>
<td>German</td>
<td>leukemia laboratory results</td>
<td>12 743</td>
<td>[30]</td>
</tr>
<tr>
<td>German</td>
<td>nephrology records</td>
<td>6 817</td>
<td>[31]</td>
</tr>
<tr>
<td>German</td>
<td>discharge reports</td>
<td>118</td>
<td>[31]</td>
</tr>
<tr>
<td>Chinese</td>
<td>medical documents</td>
<td>1100</td>
<td>[32]</td>
</tr>
</tbody>
</table>

Table 1 - Non-English corpora

The lack of appropriate lexical resources can be solved by applying unsupervised methods [33].

![Figure 1 - PubMed's publication for query: “natural language processing and (French| German| Chinese| Spanish| Japanese| Dutch| Italian| Swedish)”](image-url)
5. TOOLS FOR MEDICAL CLASSIFICATION AND ANNOTATION OF CLINICAL TEXTS

Medical terminology and classification systems are used in healthcare to facilitate interoperability among institutions and to collaborate, medical professionals, scientists, and other stakeholders. There is a justified need to integrate the various medical terminology and classification.

International Statistical Classification of Diseases and Related Health Problems (ICD) has been used since the 18th century with constant revisions and additions. It is used in over 150 countries and is available in more than 40 languages and it is under the jurisdiction of World Health Organization [34]. Classification of diseases is a system of categories that are assigned to certain diseases according to defined criteria. The International Classification of Diseases is a standard tool used in epidemiology, health management, the analysis of population health, and monitoring health problems.

SNOMED CT [35] is a structured clinical vocabulary used in any electronic health record (EHR). It is the most comprehensive and accurate clinical health terminology product in the world.

It provides that data can be shared between health and social care institutions and service providers. SNOMED CT is available in American English, English, Argentine Spanish, Danish and Swedish. Translations into French, Dutch, Lithuanian, and several others. SNOMED CT is clinical hierarchical terminology that contains medical terms and their relationships as well as synonyms, including over 320,000 terms (Figure 2, Figure 3).

Figure 2 - SNOMED CT Spanish edition 2020

Figure 3 - SNOMED CT United States edition 2020

UMLS (Unified Medical Language Systems) [36] integrates and distributes key terminology, classification and coding standards, and associated resources to promote the creation of more effective and interoperable biomedical information systems and services, including electronic health records. UMLS supports mapping between different terminologies. UMLS contains several million concepts derived from hundreds of bio (medical) dictionaries, such as ICD, SNOMED, OMIM, MeSH, GO, as well as medical abbreviations (Figure 4). It consists of three parts:

1. Metathesaurus - very large, multi-purpose, and multi-lingual vocabulary database that contains information about biomedical and health-related concepts
2. Semantic Networks - categorization and connections between all resources in metathesaurus
3. Specialized Lexicons - lexicons for biomedical and general English.

Figure 4 – UMLS concepts
Metathesaurus contains 215 different lexicons for 25 languages, of which the most are resources for English (144), followed by German, Spanish and French (Figure 5). There is interest in constantly updating these resources for different languages.

The number of thesauruses by different languages

Figure 5 – The number of thesauruses by different languages

The most commonly used UMLS products are metathesauri, followed by MetaMap [37] which is used to map concepts from metathesaurus in the text. The creation of reference corpora is crucial in the process of developing appropriate methods for solving the problems of machine translation, deidentification, drug interaction, etc. [38] [39].

The three most popular information retrieval tools are MetaMap [40], cTAKES [41] and CLAMP [42]. The common feature of these tools is to perform mapping named entity based on UMLS. MetaMap is a tool for extracting biomedical information. cTAKES is a natural language processing system for extracting data from clinical free text from electronic medical records using machine-based rules. It contains all the basic functions of NLP processing for the English language, such as tokenizer, POS tagger, named entity recognizer, negation detection, machine learning functionality, etc. The latest NLP tool for clinical text CLAMP has greater flexibility in the development of custom schemes with the possibility of their application for information retrieval. CLAMP is a Java tool, it has built-in natural language processing modules for English text. By comparing these tools in [43], it was shown that CLAMP has the best performance in terms of F1 results, and higher accuracy, and slightly lower recall compared to cTAKES and MetaMap.

Figure 6 shows an example of the application of the CLAMP tool on the example of an EHR medical report in English:

“Blood tests revealed a raised BNP. An ECG showed evidence of left-ventricular hypertrophy and echocardiography revealed grossly impaired ventricular function (ejection fraction 35%). A chest X-ray demonstrated bilateral pleural effusions, with evidence of upper lobe diversion.”

Figure 6 - CLAMP tool application example

Figure 6 shows the .xml and .txt result of mapping different medical entities in the text such as: tests, symptoms, different laboratory analyzes, and more.

6. CONCLUSION

By analyzing the existing resources for processing clinical texts in different natural languages, it can be concluded that most resources and tools are made for the English language. Great efforts are being made to create tools for other natural languages as well. The specific tools for processing the appropriate natural language are needed to be able to process clinical texts, as well as lexicons of medical terminology in the appropriate language. Some of our future goals are to create appropriate resources for the Serbian language.

7. ACKNOWLEDGEMENTS

This paper is partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia under projects III44007 and ON 174026.
REFERENCES


