Concept Map for Clinical Recommendations Data and Knowledge Structuring

Giyzel Shakhmametova[1], Nafisa Yusupova[1], Rustem Zulkarneev[2] and Yevgeniy Khudoba[1]
[1]*Computer Science & Robotics Department, Ufa State Aviation Technical University, K.Marx Str. 12, Ufa, Russia*
[2]*Faculty of General Medicine, Bashkir State Medical University, Teatralnaya Str. 2a, Ufa, Russia*
shakhgouzel@mail.ru, yussupova@ugatu.ac.ru, zurustem@mail.ru, eugchud@gmail.com

Keywords: Structuring Data and Knowledge, Unstructured Text, Clinical Recommendations, Concept Map, Production Rules.

Abstract: The article deals with the problem of structuring medical texts of clinical recommendations, which are unstructured texts. A review of existing solutions in the field of analysis of unstructured texts of both non-specialized and medical nature was carried out, shortcomings of existing developments were identified, the need for a new software solution for structuring clinical recommendations was revealed, which, in turn, is demanded in clinical decision support systems. The method of structuring data and knowledge of clinical recommendations is described, as well as the general structure of the solution, as along with the process of forming a map of concepts, including graphematic, morphological, syntactic and semantic analysis of text. In conclusion, the results of implementation in the form of concept map fragments are presented, on the basis of which further product rules are formed, which are suitable for use in knowledge bases. The method is universal and can be applied to any clinical recommendations texts.

1 INTRODUCTION

Over the past decades, the volume of stored, processed and transmitted information has increased many times in almost all areas of human activity (i.e., research, economics, and business). This has led to a significant increase in researchers' interest in data and knowledge processing techniques and algorithms.

By degree of organization, data can be conditionally divided into two categories:

- Structured data, examples of which are databases, information system logs, sensor and sensor data.
- Unstructured data, such as text data, images and videos.

According to experts [1], about 80-90% of all information used in organizations is presented in unstructured form. There is therefore a need to reduce the labour, time, financial and other resources required to process such information. The most effective way to achieve this is to bring such information into a structured form (data structuring).

Interest in methods of extracting and classifying data in unstructured texts as tools of knowledge generation has long emerged (mid-20th century [2]), but only in the last two decades have the technologies necessary for such research been developed [2]. A significant increase in interest in this field of research was caused by the advent of data processing technologies such as Text Mining and Natural Language Processing.

One of the most important areas of human activity in which it is possible to apply technologies for structuring data presented in text form is medicine [3]. In particular, the analysis of clinical documentation, i.e., medical records, survey results, operational intervention logs, etc., is of great practical importance [4] also in the context of improving health care services [5].

Among the least studied tasks in this field to date is the task of analyzing the texts of clinical recommendations. Clinical recommendations are specialized documents developed to support decision-making by a practitioner to provide appropriate medical care in a particular clinical situation. In fact, this document is the guide of the specialist in patient management, diagnosis and treatment. Clinical recommendations contain unstructured data and knowledge that guide a person skilled in the art of prescribing treatment,

examinations, and other decisions that affect the outcome of a's patient's disease [5]. In its original form, these data and knowledge are unsuitable for automated processing, and therefore clinical recommendations in medical practice are analyzed manually. If the data and knowledge of the clinical recommendations are adjusted to a structured form, it is possible to apply them in clinical decision support systems (CDSS) for the diagnosis and selection of the's patient's treatment trajectory [6].

In this article, the clinical recommendations texts structuring method is considered and examples of results obtained from texts of clinical recommendations for bronchopulmonary diseases treatment are shown. The method is universal and can be applied to other clinical recommendations texts.

2 RELATED WORKS

In the field of structuring text data for both research and application purposes, a large number of software solutions has been developed.

2.1 Solutions for Analyzing General Texts

- SAS Text Miner, an integrated component of the SAS system designed to analyze text data, provides a large set of linguistic and analytical modeling tools designed specifically to discover and extract knowledge from text information collections [7].
- GATE (General Architecture for Text Engineering) is an open source natural language processing system that uses Java component sets [8].
- STATISTICA Text Miner is an optional extension of the STATISTICA Data Miner designed to extract knowledge from unstructured texts [9].
- Natural Language Toolkit (NLTK) is a package of libraries and programs for symbolic and statistical processing of natural language written in Python programming language, containing graphical representations and sample data [10].

2.2 Solutions for Analysis of Medical Texts

The software solutions discussed above are oriented towards processing of texts of a general nature, such as news reports, for example. At the same time, it should be noted that the style of clinical texts is very different from the style of texts from other subject areas, so that their analysis requires considerable improvement of existing methods and tools for the analysis of natural language texts. Therefore, the analysis of clinical texts was identified as a separate area of research.

Research in this area has led to the development of a number of applications and platforms specializing in integrated computer language analysis of medical texts, some of which are already being used in clinics to improve the quality of medical services. Let us take a closer look at some of the most popular ones.

- UMLS (Unified Medical Language System) is a tool for the development of computer systems for the analysis of biomedical information and other types of information in the field of health care. Developed in 1986 at the National Library of Medicine (NLM) [11].
- MedLEE (Medical Language Extraction and Coding System) - a system for extracting, structuring and encoding clinical information contained in various types of medical reports (e.g., X-ray, mammography and echocardiological studies) [12].
- cTAKES (Clinical Text Analysis and Knowledge Extraction System) is an open source natural language processing system that extracts clinical information from unstructured electronic medical card texts [13].

All of the above systems have a significant disadvantage within the framework of our task: none of them have built-in support of the Russian language.

2.3 Solutions for Automatically Building Ontologies from Text Documents

A possible means of solving the problem for presenting data and knowledge of clinical recommendations is ontology, a description of the subject area presented in the form of a conceptual diagram. We looked at the most famous means of automatically generating ontologies based on text files:

- Text-To-Onto is a software solution developed by the University of Karlsruhe researchers that automatically builds ontologies based on natural language texts by identifying key

concepts in them and discovering links between them [14].

- DOG4DAG (Dresden Oncology Generator for Directed Acoustic Graphs) is a tool for automatic generation of ontologies based on natural language texts. It is presented in the form of a plugin for Protégé 4.1 and OBOEdit 2.1. This plugin allows you to use PubMed articles, web pages, or PDF documents as input. The generation of ontology in DOG4DAG is carried out by building a hierarchical model of classes connected by relationships of the form "is subclass of" [15].

As a result of the analysis on means of automatic ontologies generation (such as, for example, ASIA, Syndikate, WebKB, etc.), it was found that support for the vast majority of them is currently discontinued, many of them are unavailable, as well as none of them support Russian text processing.
In this regard, in order to solve the problem of structuring data and knowledge of clinical recommendations, it is necessary to develop an algorithm for extracting data and knowledge from Russian-language texts of medical topics.

3 PROBLEM DEFINITION

An analysis of the current state of research showed a lack of ready-made solutions in this area. Therefore, it is necessary to develop a method and algorithm for structuring data and knowledge of Russian-language texts of clinical recommendations and bringing them into a form suitable for further processing in clinical decision-making support systems, as well as their software implementation.

The software solution being developed shall have the following characteristics:

- The software decision should accept texts of clinical recommendations in Russian.
- Text processing should result in a set of rules suitable for use in clinical decision support systems (Figure1).

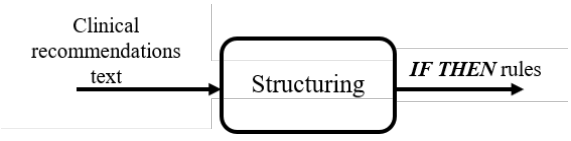

Figure 1: Task setting.

4 CLINICAL RECOMMENDATIONS DATA AND KNOWLEDGE STRUCTURING METHOD DEVELOPMENT

4.1. General Structure of a Method

In order to structure the texts on clinical recommendations, the authors proposed and developed the “concept map” object and identified the following main stages of text processing:

- Definition of keywords (performed automatically by means of specialized algorithms or manually, 2 modes).
- Mapping concepts based on keywords.
- Highlight the concepts of rules in the map and represent them in a form suitable for use in the HACT.

A diagram illustrating the above steps is shown in Figure 2.

4.2 Concept Map

The concept map is essentially a form of semantic networks and is an oriented graph whose vertices record the concepts of the subject area, and in the edges, the relations between them. Relations between concepts can be taxonomic (i.e., forming a hierarchy of concepts) and of other kinds.

A key feature of the concepts proposed by the authors of maps is the ability to display rules for links of objects based on conditional constructs. They are displayed in the diagram using dashed lines; the signature of communications for a conditional part of the rule is followed by a prefix "at", and the conclusions of the rule, by the prefix "3". An example of a concept map fragment to be developed is shown in Figure 3.

4.3 Algorithm Structure

The stage of automatic concept mapping is essential and most time-consuming, because it involves a large number of natural language processing algorithms.

The concept map development algorithm is shown in Figure 4.

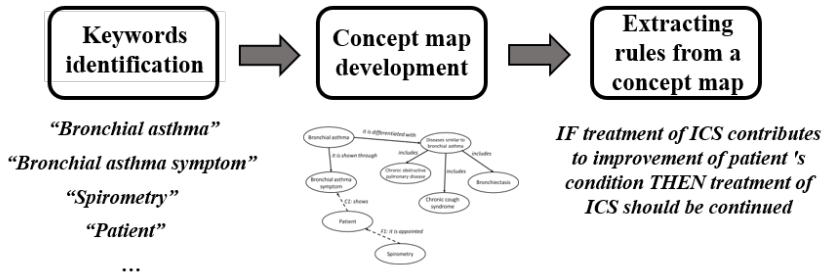

Figure 2: Stages of the process for structuring data and knowledge of clinical recommendations.

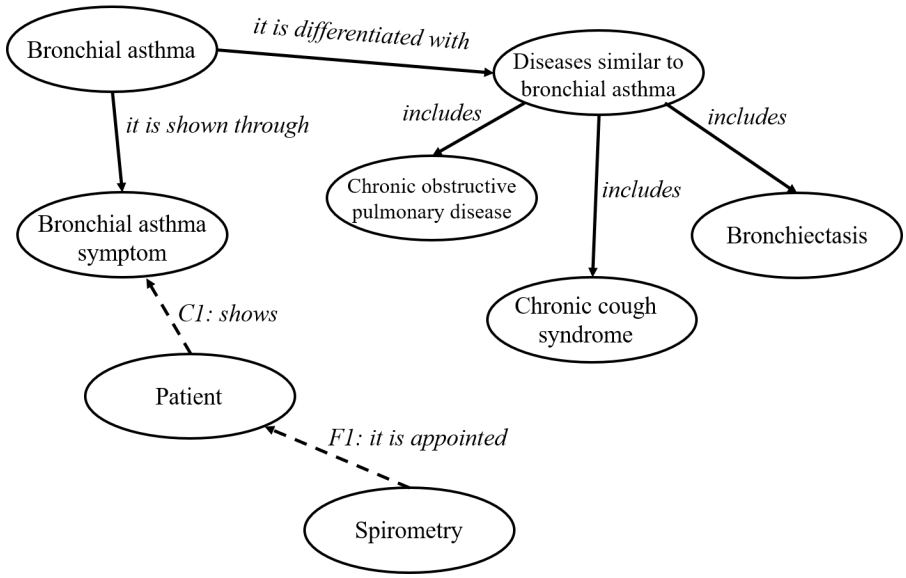

Figure 3: Concept map fragment.

In the process of constructing a concept map through this software service, the following main steps can be distinguished:

- Graphical analysis. This stage is preparatory; during this process, the text is preprocessed as required for the next steps. The tasks of this stage are to divide the source text into paragraphs, sentences, words, as well as to highlight specific words and phrases in the text (for example, proper names).
- Morphological analysis. The purpose of this step is to construct a morphological interpretation of the words of the input text. In the process of text processing, word forms are extracted from the text and subsequently normalized (lemmatized), i.e., reduction to the initial morphological form, for example, for nouns it will be a nominative case, singular, for verbs an infinitive, etc.
- Parsing. This step determines the link structure of word forms in sentences. The result of the analysis is usually presented as a so-called syntax tree (a graph of a tree structure whose nodes display word forms and whose branches are links). One of the most popular methods of conducting such analysis is the use of link grammar.
- Semantic analysis. This stage consists of highlighting semantic relations and forming semantic representation of texts. Typically, during processing, key entities (word forms or phrases) are highlighted in the text, weighted, and the strength of the links between them is counted. The result of text processing at this stage is also a graph in whose vertices concepts are placed, and in nodes - links between them. Among the methods used in practice are graph methods, varieties of

Markov random fields method, and methods of context-dependent analysis.

Building a concept map. A final step in which the graph obtained in the previous step is used to map concepts suitable for further processing.

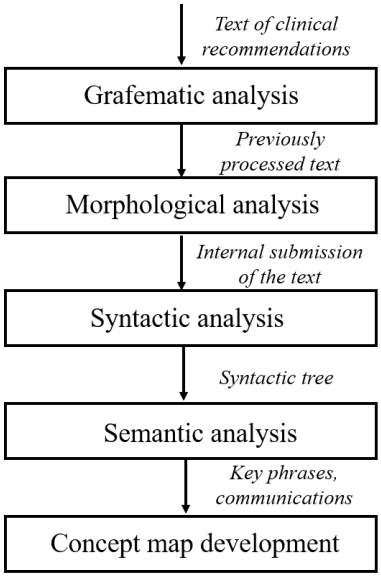

Figure 4: Concept map development algorithm.

5 REALIZATION RESULTS

The result of processing clinical recommendation text using software developed based on the methods and algorithms discussed above is a concept map, which is a mapping of key concepts found in clinical recommendation texts and the relationships between them.

Examples of the resulting concept map for clinical recommendations on the treatment of chronic obstructive pulmonary disease are shown in Figure 5.

This concept map shows the key concepts that appear in the text of clinical recommendations, as well as the relationships between them. It is necessary to separately note the presence of two conditional links represented on the map by dashed lines with prefixes (C for the conditional part of the rule and F for the final part, respectively).

Based on these conditional relationships, two rules can be generated in a format suitable for use in decision support systems:

- IF treatment of ICS (Inhaled Corticosteroids) contributes to improvement of 's patient's condition then treatment of ICS should be continued.
- IF treatment with ICS did not cause a significant change in pulmonary function THEN COPD (chronic obstructive pulmonary disease) is a probable diagnosis.

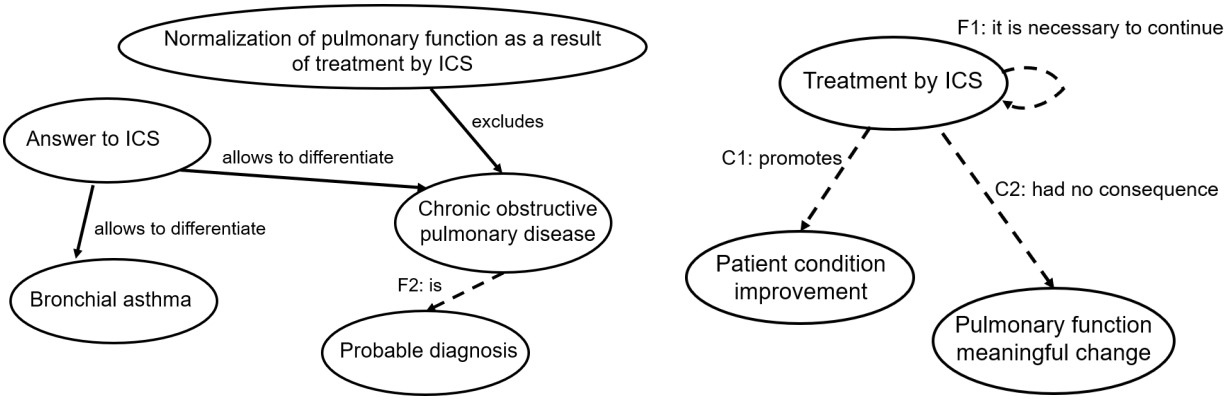

Figure 5: Fragments of the realization results.

6 CONCLUSIONS

An analysis of existing solutions for the task of automatic extraction of data and knowledge from the texts of clinical recommendations has shown that none of the currently available software is suitable for the task in question. In this regard, it was concluded that a new method of structuring data and knowledge of clinical recommendations should be developed and implemented as a software solution.

The proposed method distinguishes the use of a new means for presenting data and knowledge in a structured form called a "concept map"; i.e., it is possible to represent the relationships between the concepts containing the conditional and the final part. Since one of the key stages of the developed method for automatic extraction of data and knowledge from clinical recommendations is generation of product rules from obtained maps of concepts, it is possible to further apply these rules in knowledge bases of clinical decision support systems.

ACKNOWLEDGMENTS

The reported study was funded by RFBR according to the research projects № 19-07-00780, 19-07-00709.

REFERENCES

[1] R. Grishman, 'Twenty-five years of information extraction". Natural Language Engineering, vol. 25(6), 2019, pp. 677-692, doi: 10.1017/ S1351324919000512.

[2] S. Grimes, "A Brief History of Text Analytics". Eye Network. Retrieved, June 2016.

[3] B. Presannan, N. Ramasubramanian and A.S. Vijayan, "Disease risk prediction from clinical texts", 2020, doi:10.1007/978-981-32-9515-5_30.

[4] F. Dhombres, J. Charlet and Section Editors for the IMIA Yearbook Section on Knowledge Representation and Management. Formal medical knowledge representation supports deep learning algorithms, bioinformatics pipelines, genomics data analysis, and big data processes. Yearbook of Medical Informatics, vol. 28 (1), 2019, pp. 152-155, doi: 10.1055/s-0039-1677933.

[5] B. Séroussi, L.F. Soualmia and J.H. Holmes, Transforming data into knowledge: How to improve the efficiency of clinical care? Yearbook of Medical Informatics, vol. 26(1), 2017, pp. 4-6, doi:10.1055/s-0038-1637768.

[6] C. Combi and G. Pozzi. "Clinical information systems and artificial intelligence: Recent research trends". Yearbook of Medical Informatics, vol. 28(1), 2019, pp. 83-94, doi:10.1055/s-0039-16779156.

[7] Text Mining Software, SAS Text Miner | SAS. Renewal date: 18.03.2019. [Online]. Available: https://www.sas.com/en_us/ software/text-miner.html.

[8] General Architecture for Text Engineering. Renewal date: 18.03.2019. [Online]. Available: https://gate.ac.uk.

[9] STATISTICA Text Miner. Renewal date: 18.11.2019. [Online]. Available: http://statsoft.ru/products/STATISTICA_Data_Miner /STATISTICA_Text_Miner.

[10] Natural Language Toolkit NLTK 3.4 documentation. Renewal date: 18.03.2019. [Online]. Available: https://www.nltk.org.

[11] Unified Medical Language System (UMLS). Renewal date: 18.11.2019. [Online]. Available: https://www.nlm.nih.gov/ research/umls.

[12] MedLEE | MedLingMap. Renewal date: 18.11.2019. [Online]. Available: http://www.medlingmap.org/ taxonomy/term/80.

[13] Apache cTAKES - clinical Text Analysis Knowledge Extraction System. Renewal date: 18.11.2019. [Online]. Available: http://ctakes.apache.org.

[14] A. Maedche and S. Staab, "The TEXT-TO-ONTO Ontology Learning Environment" (PDF). ResearchGate, July 2000.

[15] Dresden Ontology Generator for Directed Acyclic Graphs (DOG4DAG). Renewal date: 30.10.2019. [Online]. Available: http://www.biotec.tu-dresden.de/ research/schroeder/dog4dag.