Complex Term Identification for Ukrainian Medical Texts

Olga Cherednichenko [0000-0002-9391-5220], NadiiaBabkova [0000-0002-2200-7794]

and Olga Kanishcheva [0000-0002-9035-1765]

National Technical University "Kharkiv Polytechnic Institute",
2, Kyrpychova str., 61002 Kharkiv, Ukraine
olha.cherednichenko@gmail.com, Nadiia.Babkova@khpi.edu.ua,
kanichshevaolga@gmail.com

Abstract. The medical texts are very difficult in understanding when we have not only complex words in general meaning, but a lot of special terms and notions. It causes difficulties in understanding texts in medicine domain. Natural language processing engages people all over the world to apply statistics, machine learning, deep learning, and linguistics in order to solve those tasks. Linguistically complex tasks, such as the medical text understanding, are the most challenging because they require linguistic intuition. In this paper, we study how linguistic approach can be applied to solve the problem of identification of complex words. In order to study medical texts simplification, we try to analyze medical unified protocols as the case and test nlp approach for medical words identification. A dataset of different medical protocols from official resources is developed. The features of special medical words are studied.

Keywords: Term Identification, Medicine Text, Text Simplification, Explanation, NLP, Ukrainian Texts.

1 Introduction

Medicine, biology, pharmacology and other neighboring areas are well-known to be overloaded with complex terms and notions. Medical texts contain many borrowed words that came from Latin, for instance. Moreover, terms may have multiple synonyms which make the texts even more complicated. Non-expert readers may get additional use of simplified medical texts in several situations. Firstly, when a person doesn't have a medical education, a simplified text may become useful to get out an idea what some medical instructions actually mean. For example, when a medical prescription defines to make some tests, the test name can be quite complex (like, biochemical blood assay) while for the patient it will mean just a blood test that requires some particular preparation before it. Secondly, a person may wish to obtain a more comprehensive explanation of his/her diagnosis written in the medical assessment report. A simplified text, in this case, may help to understand the character of disorder but not the accurate and detailed diagnosis. Thirdly, after a patient gets the subscription

from the physician that contains many complex medical terms, it may lead to confusions. Therefore, text simplification can help a person to follow the doctor's instructions and get some clear explanations what particular treatment methods mean.

Medical information is available on the web in the information systems of medical establishments, healthcare portals, medical libraries, etc. More and more non-expert readers would like to use this information. The digital character and availability of this data together with a broad range of potential readers induce the development of an information system for medical texts simplification.

Text simplification problem arises when the initial text has to be modified in order to make it more readable and understandable for the audience[1]. There can be different reasons why the initial text looks inappropriate for usage [2]. The most probable cases are the result of the text complexity itself or the peculiarities of the audience that wants to use the text. Anyway, simplification of the texts requires modern methods of language processing and data analysis [3, 4].The difficulty of the text's syntactic and lexical structure may lead to many inconveniences for the readers. For example, the text might be too complex for the people with some specific disabilities and disorders which make it hard to perceive the information. Other categories of potential readers who may face problems while using complex texts include children, low literacy people, language learners, etc. For all of them, a simplified text would be a good solution to get the idea and use of the textual information.

Another area where the problem of texts simplification becomes quite essential is machine processing of textual data. The original texts may be complex enough for Natural Language Processing (NLP) techniques. Therefore, in order to solve some problems of language processing, it would be convenient to apply NLP algorithms to the text that has already been simplified previously. Such problems include information retrieval and parsing, information summarization and annotation, machine translation, etc.

This paper represents the empirical study of medical information simplification in order to increase the readability and comprehension of original medical texts.

2 Related Works

The process of decreasing the linguistic complexity of a text, and retaining the original information and meaning is the problem of text simplification. Text simplification can be used for many purposes: second language learners, preprocessing in pipelines and assistive technology, to automatically extract of data from doctor's notes, laboratory results and other medical documents [5-7]. In order to properly represent textual information into computable forms for a certain task like classification, clustering, sentiment analysis, recommendation and information retrieval different ways of text simplification are used.

Nowadays automatic lexical simplification systems either do not have sufficient coverage (supervised approaches), or they only perform one-to-one word substitutions and thus cannot simplify longer lexical phrases, and they do not perform any kind of word reordering [8]. In the paper [9], an initial dataset for automated text simplification using

a refined set of operationalized guidelines for manual simplification were create and methodology for expanding the dataset was develop. Adaptation of statistical machine translation to perform text simplification, taking advantage of large-scale paraphrases learned from bilingual texts and a small number of manual simplifications with multiple references is presented in [10]. Semantic term weighting which considers term meanings is significant for specific applications of machine learning [4].

The use of text simplification as a pre-processing step for statistical machine translation of grammatically complex under-resourced languages is explored [11]. The experiments on English-to-Serbian translation show that this approach can improve grammaticality of the translation output and reduce technical post-editing effort (number of post-edit operations). Simplification can be applied on lexical, syntactic, and discourse level [12]. Some lexical plugins [13] allow the use of different synonyms in order to avoid repetition and in the case of the syntactic simplification, the user to see all the conjunctions in the text and to separate complex sentences with a few simple clicks.

3 Methods

Medical texts include drug packages, medical records, fact sheets, medical reference books, and training materials, certificates, etc. To solve the problem of the simplification of a medical text, it is first necessary to single out the features of such texts. In this study, we rely on the texts of medical clinical protocols. In order to accelerate the development and implementation of the state standards in the field of health, the Ministry of Health of Ukraine approves medical and technological documents on the basis of evidence-based medicine. Such documents include a unified clinical protocol for medical care, as well as an adapted clinical trial that based on evidence. Depending on the disease, the plan of treatment and preventive measures may differ, which is also prescribed in the legislation in the local protocols of prevention and treatment.

In the modern conditions of Ukraine, which is actively integrating into the European and world community, health care reforms have been adopted. These reforms provide a legal basis for evidence-based medicine. A universal clinical protocol provides the foundation for the functioning of medical institutions and private doctors. Therefore, in this work, it is the texts of medical protocols that are used to construct the corpus of medical texts for the analysis of linguistic complexity.

At the first stage, the typical medical protocol is considered. Let us consider, as an example, a protocol for the prevention of cardiovascular disease [14]. According to the order of the Ministry of Health of Ukraine, this protocol defines signs and criteria for the diagnosis of the disease; the conditions in which the medical aid should be provided; a diagnostic program consisting of compulsory and additional research; medical program; recommendations for the further prevention of medical care.

Clinical protocols, as well as other guidance documents, are publicly available on the Internet [14]. They are a guide for medical practitioners, for the administration of health facilities, and also a source of information for patients. All protocols before approval pass multiple collective expertise. The data that is entered in the protocol meets the medical standards of the national and international levels.

The analysis shows that all protocols have a common structure: introduction, abbreviations, passport part, general part, main part, description of the stages of medical care, resource support for the implementation of the protocol, indicators of the quality of medical care, a list of references and appendixes.

The main idea of our research is the simplification of the medical text depends on the complexity of this text and the stakeholder, who studies this text. So, for patients, such parts of the protocol as a passport of the protocol, or a list of references, can be omitted. For patients, those parts of the protocol that describe the symptoms of the disease, the epidemiology, the necessary actions of the doctor and, especially, the recommendations are of the greatest interest. It should also be noted that all medical records are provided in the state language. As a result, the medical text is replete with not only Latin special terms, but also complex medical words in the Ukrainian language. Analysis of the Ukrainian text in terms of linguistics is a daunting task. In this case, the problem is complicated by the huge amount of medical terminology. At the same time, the text also contains words from the subject area, which do not require simplification. The proposed pipeline (Fig. 1) is quite general for such kind of tasks.

We would like to underline the step of complex word identification that is based on feature extraction. We presuppose to use morphological templates to identify complex medical words. The aim of this empirical study is to realize the approach to in complex word identification process for Ukrainian medical texts.

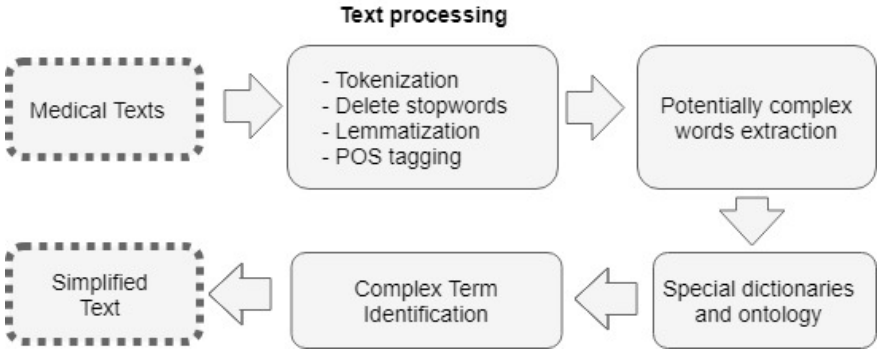


Fig. 1. A pipeline for medicine protocol processing

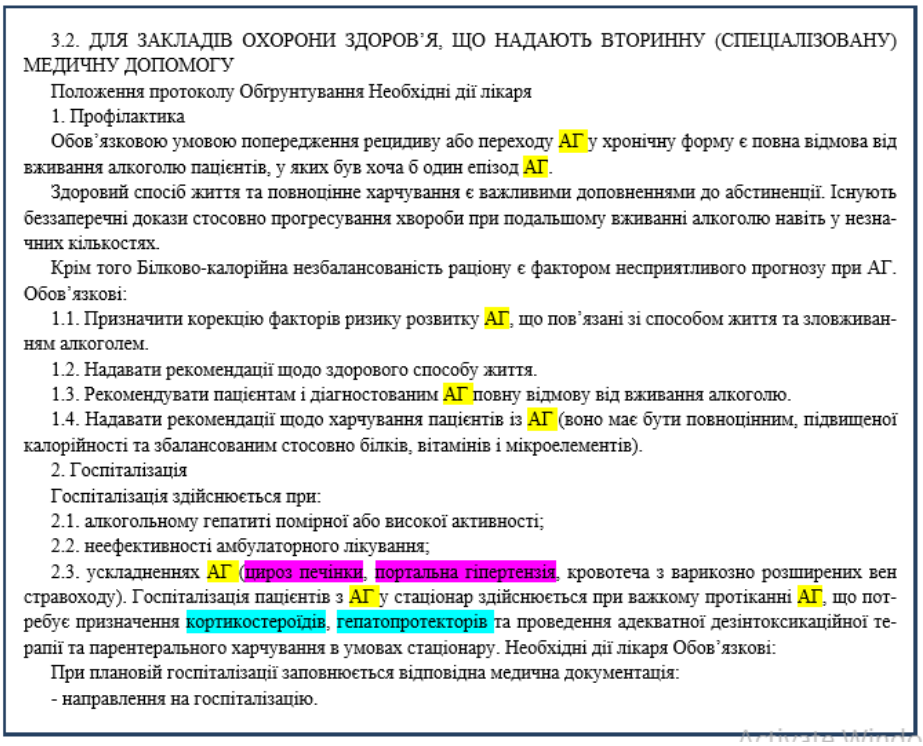
4 Experiments and Results

4.1 Dataset

We have twenty existing clinical protocols [14] for our experiments, which contain more than 2,500 thousand words. Protocols were taken from the site "Register of medical-technological documents" (http://mtd.dec.gov.ua) and they have several directions, such as dermatovenereology, genetics, gastroenterology, dermatology, allergology, and hematology. The protocols' texts have several obligatory parts, however, inside the

paragraphs, they are weakly structured. The example of the original text is presented in Fig. 2.

In Fig. 2 we singled out such parts of the document: yellow color is abbreviations (АГ); blue color is complex medical words (*кортикостероїдів/ corticosteroids, гепатопротекторів/ hepatoprotectors*) and pink color is special medical terms (*цироз печінки/ liver cirrhosis, портальна гіпертензія/ portal hypertension*).



3.2. ДЛЯ ЗАКЛАДІВ ОХОРОНИ ЗДОРОВ'Я, ЩО НАДАЮТЬ ВТОРИННУ (СПЕЦІАЛІЗОВАНУ) МЕДИЧНУ ДОПОМОГУ

Положення протоколу Обгрунтування Необхідні дії лікаря

1. Профілактика

Обов'язковою умовою попередження рецидиву або переходу АГ у хронічну форму є повна відмова від вживання алкоголю пацієнтів, у яких був хоча б один епізод АГ.

Здоровий спосіб життя та повноцінне харчування є важливими доповненнями до абстиненції. Існують беззаперечні докази стосовно прогресування хвороби при подальшому вживанні алкоголю навіть у незначних кількостях.

Крім того Білково-калорійна незбалансованість раціону є фактором несприятливого прогнозу при АГ. Обов'язкові:

1.1. Призначити корекцію факторів ризику розвитку АГ, що пов'язані зі способом життя та зловживанням алкоголем.

1.2. Надавати рекомендації щодо здорового способу життя.

1.3. Рекомендувати пацієнтам і діагностованим АГ повну відмову від вживання алкоголю.

1.4. Надавати рекомендації щодо харчування пацієнтів із АГ (воно має бути повноцінним, підвищеної калорійності та збалансованим стосовно білків, вітамінів і мікроелементів).

2. Госпіталізація

Госпіталізація здійснюється при:

2.1. алкогольному гепатиті помірної або високої активності;

2.2. неефективності амбулаторного лікування;

2.3. ускладненнях АГ (цироз печінки, портальна гіпертензія, кровотеча з варикозно розширених вен стравоходу). Госпіталізація пацієнтів з АГ у стаціонар здійснюється при важкому протіканні АГ, що потребує призначення кортикостероїдів, гепатопротекторів та проведення адекватної дезінтоксикаційної терапії та парентерального харчування в умовах стаціонару. Необхідні дії лікаря Обов'язкові:

При плановій госпіталізації заповнюється відповідна медична документація:

- направлення на госпіталізацію.

Fig. 2. Example of a paragraph of unified clinical protocol of the primary, secondary (specialized) medical aid from alcohol hepatitis

4.2 Our Results

We used a Pymorphy2 program as the morphological analyzer (https://github.com/kmike/pymorphy2) and a list of stopwords to preprocess the texts. Our list of stopwords was created from different resources and contains such parts of speech as pronouns, adjectives, prepositions, exclamations, suffixes, a combination of letters etc. On the first step, we tried to use simple frequency for term identification. Examples of the most and least frequently used words are presented in Table 1 and Table 2, respectively. These results were very poor, the necessary words were in different parts of the term list.

Table 1. Fragment list of the most frequently used words.

Word (eng/ukr)	Frequency
лікування/treatment	1427
пацієнт/patient	1084
МОЗ/Ministry of Health	726
Україна/Ukraine	717
здоров'я/health	476
індикатор/indicator	473
протокол/protocol	440
лікарь/doctor	421
дитина/child	379
розвиток/development	349
проведення/carrying out	334
обстеження/examination	329

In Table 2 contains not only medical special terms such as *сероконверсія* (*seroconversion*), *гломерулонефрит* (*glomerulonephritis*) etc. but and words of general medical vocabulary *вагітність* (*pregnancy*), *спектроскопія* (*spectroscopy*). Thus, we cannot use simple frequency for finding of special medical terms.

Table 2. Fragmentlist of the least frequently used words.

Word (eng/ukr)	Frequency
сероконверсія/ seroconversion	4
гломерулонефрит/ glomerulonephritis	4
опіоїд/ opioid	4
моксифлоксацин/ moxifloxacin	4
сертралін/ Sertraline	4
остеопенія/ osteopenia	4
вагітність/ pregnancy	4
спектроскопія/ spectroscopy	4
пентоксифілін/ Pentoxifylline	4
гептагідрат/ heptahydrate	4

As the next step, we selected nouns from the general frequency lexicons. This allowed us to find in the text's terms, which have the following characteristic structure: *noun + noun* or *adjective + noun* (*adjective + adjective + noun*). Lexicons of compound terms for *noun + noun* construct are presented in Table 3 and Table 4.

Table 3. List of compound terms (*noun + noun*).

Word (eng/ukr)	Frequency
охорона здоров'я/ health care	875
кваліфікаціяхвороб/ qualification of diseases	782
дитинаінвалід/ child is disabled	218
інфаркт міокарда/ myocardial infarction	203
забезпечення якості/ quality assurance	194
синдром Дауна/ Down syndrome	167
стан здоров'я/ health status	154
хвороба гоше/ Gaucher's disease	48
емфіземалегень/ emphysema of the lungs	29
цироз печінки/ cirrhosis	14

Table 4. Part of the list of the least frequently used words.

Word (eng/ukr)	Frequency
практика сімейних/ family practice	43
контрольний рівень/ control level	42
хірургічне втручання/ surgical intervention	35
підвищений ризик/ high risk	35
шлунково-кишковийтракт/ gastrointestinal tract	33
хронічний кашель/ chronic cough	27
протозойна інфекція/ protozoal infection	26
бронхіальна астма/ bronchial asthma	18
вірусний гепатит/ viral hepatitis	18
церебральний параліч/ cerebral palsy	12

Accordingly, we received the statistics about our lexical templates. This information is presented in Table 5. The total number of nouns is 5,973 in our dataset documents.

Table 5. The statistics of lexical templates.

Lexical template	Number
noun	1,322
noun + noun	127
adjective + noun	283
adjective + adjective + noun	46

We try to change our templates to identify complex medical terms. The issue is the quality and quantity of data in the templates. We use our testing data set to evaluate the obtained results. The precision, recall and F-measure are presented in table 6.

Table 6. The evaluation values.

Lexical template	Precision	Recall	F_1-measure
noun	0,53	0,48	0,5
noun + noun	0,5	0,43	0,46
adjective + noun	0,6	0,67	0,63

Quality analysis of experiment showed unsatisfied results. The main reason for such a point is that we analyzed data set which is on Ukrainian. Morphological analyzer works not properly enough on Ukrainian texts. The list of stop-word was changed but it didn't cause better results in special words searching. The quality results claim that templates occur often in texts, but for their automatic extraction morphological analyzer is needed. Besides that, it's reasonable to investigate sentences structure in medical tests. Issues mentioned above are topics for future research.

5 Conclusion and Future Works

In our work, analysis of the frequency dictionary showed that specific and complex terms are used less frequently, which allowed us to discard the words with the highest frequency of occurrence. Such wise, the results of our experiments show that we can distinguish two main groups that require simplification:

1) Specific words (for example, *glucocorticosteroid*), which can be replaced by more simple "spoken" synonyms;

2) Compound terms that can be simplified by clarifying their meaning. It will influence the syntactic structure of a sentence.

For the first group, it will be advisable to use a dictionary of synonyms. To work with the second group, it is necessary to create the special terminology ontology or another lexical resource.

The experiment shows that special medical texts, such as protocols, are written according to a specific pattern, with the result that words and phrases such as "*treatment*", "*patient*" or "*Ukraine*" are found hundreds of times more often than other nouns. Among the words with an average frequency (about 100 in our experiment) are found as well as recognized words (*professor* − 91, *document* − 91), and complex (*clinical examination* − 91, *syndrome* − 178), as well as diagnoses (*hepatitis* − 123, *tuberculosis* − 141). It should be noted that among the words with a low frequency (less than 60) the share of special medical terms is 0.87. This confirms the hypothesis that the frequency of the word in the corpus of medical texts is a sign of its complexity. At the same time, words found in this way require additional research by other methods.

References

1. Siddharthan, A.: A survey of research on text simplification. In: International Journal of Applied Linguistics, Peeters Publishers, Belgium (2014) doi.org/10.1075/itl.165.2.06sid

2. Shardlow, M.: A Survey of Automated Text Simplification. In: International Journal of Advanced Computer Science and Applications,pp.58-70(2014)
3. Falkenjack, J.etal.: Services for Text Simplification and Analysis. In: Proceedings of the 21st Nordic Conference of Computational Linguistics, pp.309-313, Gothenburg, Sweden, 23-24 May (2017)
4. Matsuo, R., Tu Bao Ho: Semantic Term Weighting for Clinical Texts. In: Expert Systems With Applications (2018) doi.org/10.1016/j.eswa.2018.08.028.
5. Popolov, D., Barr, J. R.: Units of meaning' in medical documents. In: IEEE International Conference on Semantic Computing, pp. 320-323 (2014) doi.org/10.1109/ICSC.2014.62
6. Mukherjee,P. etal.: NegAIT:A new parser for medical text simplification using morphological, sentential and double negation. In: Journal of Biomedical Informatics 69, pp.55-62 (2017)
7. Sridevi, M., Arunkumar, B.R.: Information Extraction from Clinical Text using NL Pand Machine Learning: Issues and Opportunities. In: International Journal of Computer Applications, pp.11-16 (2016)
8. Stajner, S.,Saggion, H.,Ponzetto, S. P.: Improving lexical coverage of text simplification systems for Spanish. In: Expert Systems with Applications, vol.118, pp.80-91 (2018)
9. Djamasbi, S.: Improving Manual and Automated Text Simplification (2017)
10. Xu, W.etal.: Optimizing Statistical Machine Translation for Text Simplification. In: Transactions of the Association for Computational Linguistics, vol.4, pp.401-415 (2016).
11. Baltic,J.: Can Text Simplification Help Machine Translltion In: Modern Computing, vol. 4, No.2, pp. 230-242 (2016)
12. Stajner, S.,Glavas, G.: Leveraging event-based semantics for automated text simplification. In: Expert Systems With Applications 82, pp.383-395 (2017)
13. Hervas,R. etal.: Integration of lexical and syntactic simplification capabilities in a text editor. In: Procedia Computer Science, 27, pp.94-103 (2014)
14. Registry of medical and technological documents http://mtd.dec.gov.ua/index.php/uk/
15. Abrahamsson,E. etal.: Medical text simplification using synonym replacement: Adapting assessment of word difficulty to acompounding language. In: Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) @EACL 2014, pp.57-65 (2014)
16. Jackson,R. G., PatelR., Jayatilleke N.etal.: Natural language processing to extract symptoms of severe mental illness from clinicaltext: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE). In: project BMJOpen 2017 (2016) doi.org/10.1136/bmjopen-2016-012012
17. Nojavan,F. etal.: Explanation of "taghtirolbol" in traditional medical texts Which one Dribbling or Pollakiurea. In: Journal of Islamic and Iranian Traditional Medicine, vol.6, No. 2 (2015)
18. Chen,J. etal.: A Natural Language Processing System That Links Medical Terms in Electronic Health Record Notes to Lay Definitions: System Development Using Physician Reviews. In: JMedInternetRes 2018 (2018) doi.org/10.2196/jmir.8669
19. Salah, Ait-Mokhtaretal.: A Framework to Generate Sets of Terms from Large Scale Medical Vocabularies for Natural Language Processing. In:European FP7-ICTEURECA project:http://eurecaproject.eu/ (2015)