Collection and Processing of a Medical Corpus in Ukrainian

Olga Cherednichenko [1][0000-0002-9391-5220], Olga Kanishcheva [1][0000-0002-9035-1765], Olena Yakovleva [2][0000-0002-6129-6146], Denis Arkatov [1][0000-0003-0162-059X]

[1] National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine
[2] Kharkiv National University of Radio Electronics, Kharkiv, Ukraine
olha.cherednichenko@gmail.com, kanichshevaolga@gmail.com,
olena.yakovleva@nure.ua, denarkatov@gmail.com

Abstract. The text corpora are the basis of natural language studying. We describe the structure of a Ukrainian-language corpus (UKRMED), which contains a variety of medical text genres (Clinical protocols, Blogs, and Wikipedia). The paper shows the process of collecting, creating and processing a corpus of medical data in Ukrainian. We represent our own framework for creating a text corpus. The medical domain and text simplification are chosen as corpus directions. The authors gave statistical characteristics of the corpus, an analysis of the morphological parts of speech is provided. Frequency lemmas for this medical corps are analyzed. The UKRMED corpus can be used for solving the task of natural language simplification.

Keywords: Medicine Corpus, Corpus Linguistic, Ukrainian, Text Collection.

1 Introduction

The text corpora are the basis of natural language studying, namely the medicine corpora is selected as a domain for research. It is important for everybody to understand the healthcare records and doctor recommendations regarding their health or the health of their family. On the other hand, the Internet contains a huge amount of medical information from descriptions of diseases and symptoms to recommendations for prevention and treatment, including descriptions of medicine. It is noticed that medical texts often contain many special terms and abbreviations, which complicate their understanding by ordinary people. Solving the problem of medical text simplification [1], we were faced with the lack of suitable data sets for conducting experiments with the Ukrainian language. Thus, the goal of the given study is to create a Ukrainian text corpus in the medical domain in order to study the challenges of special text simplification.

In general, meaning, a corpus is a collection of interconnected documents in a natural language. Usually studying a natural language model begins using a generalized or special corpus. For example, there are a lot of examples of scientific papers based on ready-made data sets, as a rule the text corpus from Wikipedia is used. However, the most successful language models are often highly specialized for a particular domain. It leads to the task of creating and developing a specialized text corpus. In the given paper we investigate the issues of text corpus creation in the medical domain as well as particularities of the Ukrainian language in the medical field.

We can underline some features of Ukrainian medical texts. Medical texts contain many borrowed words that came from Latin, for instance. Sometimes, terms and notions may have a variety of synonyms, which make the text even more complicated. It is really difficult to percept such specific domain texts by ordinary people. We suggest that non-expert readers may query additional use of simplified medical texts in the following cases. Firstly, in the case a person does not have a medical education, a simplified text may become useful to find out an idea of what some medical prescriptions actually mean. For example, when a medical instruction defines to make some examinations, the test name can be quite complex (like, biochemical blood assay) while for the patient it will mean just a blood-based test that requires some particular preparation before it. Secondly, a person may wish to know a more comprehensive explanation of his/her diagnosis written in the medical records. A simplified text, in this case, may help to specify the issue as well as clarify the features, possible consequences and understand the recommendations. Thirdly, a patient should have a possibility to distinguish truth and fake information from the Internet sources. Therefore, medical text simplification can help people to follow the doctor's instructions and obtain clear explanations of meanings of particular treatment methods or prescriptions.

In addition, the area where the task of text simplification becomes quite important is text automated processing. The original texts may be quite complex for Natural Language Processing (NLP) techniques. It requires the pre-simplified text would be used to applying NLP algorithms for language processing. There are information retrieval and parsing, information summarization, and annotation, machine translation, etc. that may be included as examples of such problems.

This paper represents the empirical study towards medical text corpus creation in order to investigate the readability and comprehension of original medical texts in Ukrainian.

2 Background and Related Work

NLP is a field of computer science and computational linguistics that study the peculiarities of communication between software and humans under the meaning of language. Many areas in NLP relate to natural language understanding and the opportunity to derive sense from natural language input [1, 3, 6, 13]. One of the traditional NLP tasks is the creation and development of text corpus. Many authors deal with the problems of corpus linguistics [2, 4, 7, 14].

The medical texts are very difficult in understanding due to they contain not only complex words in the general meaning but a lot of special terms and notions from the medical domain. It causes difficulties in perception texts and inconvenience in reading. Natural Language Processing techniques suggest applying 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 due to their specific and they require peculiar models and methods.

We started our research with the task described in [1], where we investigated how a linguistic approach can be applied to solve the problem of the identification of complex words. The discourse of our research is Ukrainian medical texts. In order to study medical text simplification, we analyze medical unified protocols and test the NLP approach for medical word identification [1].

The medicine as a domain for NLP researchers is not new. Some articles are devoted to the creation of text resources to specify medical discourse. For instance, user opinion mining is considered in the paper [5]. Opinion mining in the medical field has not been studied deep. Text corpus in the medical domain, such as polarized lexicons, is presented in [5] for this task.

A method for the comparison of the simplified and original texts is presented in [6]. The goal of the paper is to level off the parallel text corpus, which consists of news in Spanish with their simplified sample. In this paper, the provided algorithm is used for the creation of a corpus for the study of text simplification.

The paper [2] presents a work in progress to create an annotated text corpus for Latvian. An important aspect that considered in [2] is the variety and balance of the corpus in terms of genres, authors and lexical units. The paper [3] is considered issues of corpus creation as well. Presented by authors the framework provides several built-in NLP tools to automatically preprocess texts and is highly customizable [3].

The development of a doctor-patient dialogue corpus to support a speech-to-speech machine translation effort for English-Persian medical dialogues is described in the paper [4]. The described corpus was developed by recording and transcribing dialogues in English, and then translated into Persian. The authors highlight the benefits and drawbacks of creating a corpus in this way. Benefits include the ability to customize the corpus in a way that would be infeasible for actual doctor-patient data and avoidance of privacy and legal issues, while drawbacks include the fact that the Persian does not originate as speech, but as text translation of English speech [4].

Another side of corpus linguistics concerns low resource languages. Building language corpora for low resource languages is challenging because of limited digitized texts [7]. Language corpora are needed for building information retrieval services such as search and translation and to support further online content creation. A novel solution to source relevant multilingual content is proposed in [7]. The way of gathering data is performed by crowdsourcing translations via an online competitive game where participants would be paid for their contributions [7].

Some researchers outline the significance of combining results from different study areas. Therefore, as it is shown in the paper [8] agglomerating results from studies of individual biological components may produce biomedical discovery and the promise of therapeutic development. Such knowledge integration could be facilitated by automated text mining. The paper [8] notices that the creation of appropriate datasets is hampered by the absence of a resource for launching a distributed annotation effort, as well as by the lack of a standardized annotation schema. The annotation schema and the corresponded tool are proposed in [8]. Those proposals can be widely adopted so that the resulting annotated corpora from a multitude of disease studies are assembled into a unified benchmark dataset.

The paper [9] is devoted to studying some cultural peculiarities of rendering English texts into Ukrainian. The authors of the study [9] emphasize the importance of the experience exchange in order to strengthen the ties with economically developed countries, as well as to improve the level of professional and ethical training of current and future physicians. The studied type of text combines the features of both medical and moral-ethical discourses, thus causing some difficulties in the adequate translation from English into Ukrainian.

The purpose of the study [10] is to study the readiness of Ukrainian psychiatrists to introduce a new classification reflecting the latest trends in the integration of mod-ern neurobiological research into clinical practice. The new terminology and new classification at a certain domain are studied based on survey results in European countries and Ukraine. The authors point out the relevance and precision of the terms used in the medical domain [10]. It highlights the relevance of our research direction.

An analysis of the references shows that little attention is paid to the creation of a linguistic corps, especially the Ukrainian language and in the medical domain. However, many authors note the importance of studying medical texts. Therefore, we propose to focus on the process of text corpus creation in this paper. The aim of the given work is the medical text corpus in Ukrainian as a reference data set for further study.

3 Collection and Processing of Texts

3.1 Corpus Creation

Creating a quality text corpus is a challenge due to the significant influence of initial data on processing results. There are no defaulted technics for the creation of text data set in the theory of corpus linguistics. We suggest our own way to gather texts in the field of medicine in order to perform the medical text corpus. The main requirement of the corpus is the ability to provide data for language issues study. To create a medical corpus in Ukrainian, this study proposes the following pipeline (Fig. 1). Wikipedia texts, as well as texts from Social Networks, blogs, digital libraries, as well as official websites of medical clinics and the ministry of health, are considered as data sources. The first step is identifying the data sources and estimation of Ukrainian medical text availability on the chosen data sources.

Data is collected under the assumption that three categories of text complexity can be distinguished. We give out three classes of text complexity such as complex, moderate and simple texts. We apply our own perception and attitude of medical texts to divide them into three categories. One of the most important requests on the second pipeline step is collecting the same amount of data in each group. In addition, we try to use semantically close sources to collect text data. At the pre-processing stage, the text is cleared of unnecessary characters, as well as conversion to encoding UTF-8. The third step includes such common preprocessing actions as tokenization, normalization and makes description with a set of statistical indicators.

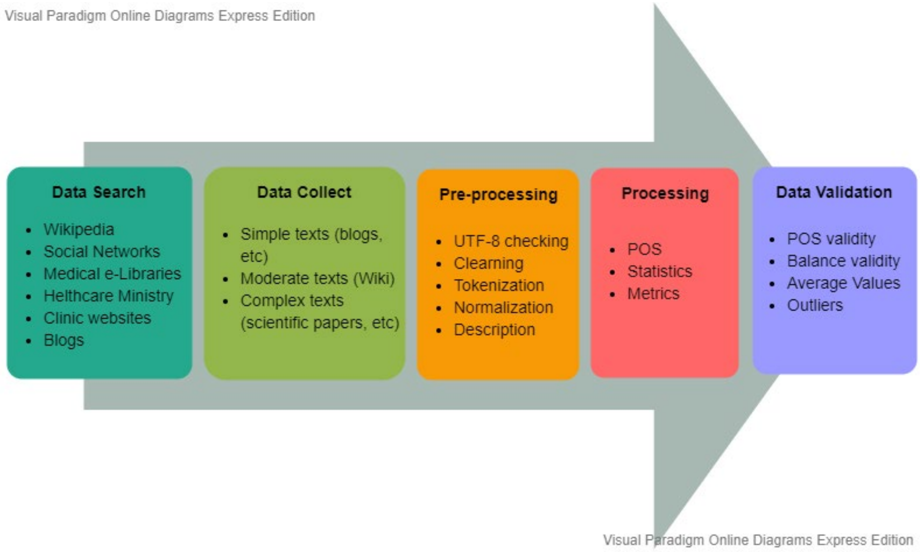

Fig. 1. Text corpus creation pipeline

The simplest and most common way to organize a text corpus for managing it is to store documents in a file system on the disk. By organizing the placement of documents in the corpus into subdirectories, you can ensure their classification and meaningful separation according to available meta-information, such as dates or sources. Thanks to the storage of each document in a separate file, corpus reading tools can quickly search for different subsets of documents, the processing of which can be carried out in parallel when each process receives its own text, different from the other subset of documents. We use this way of storage in the given research.

The pre-validation of the text data is carried out manually and some texts can be returned to the pre-processing step. We perform the crowdsourcing technic for manual data validation. At the processing stage, we suggest to determine the parts of speech and obtain the statistical descriptors of the received texts. The final step is the validation of collected text data. First of all, we check the POS tags. The balance of the text corpus is paid attention, as well as analysis of statistical characteristics, should be done. Prepared and marked up texts are placed in the repository.

3.2 Corpus Description

Our corpus (UKRMED), the UKRainian MEDicine text corpus, combines three medical text genres with a focus on clinical protocols (*"Complex texts"*), medicine forums (*"Simple texts"*) and texts from Wikipedia (*"Moderate texts"*). The clinical protocols and other complex texts are taken from the official website of the Ministry of Health of Ukraine (https://guidelines.moz.gov.ua), dissertations and scientific papers (http://idvamnu.com.ua/journal). The texts from the *"Simple text"* category are taken from such sites as KinesisLife (https://kinesislife.ua), hospital sites (https://gynecology.kyiv.ua), Dr. Komarovsky (https://komarovskiy.net), etc. The data from *"Moderate texts"* category is taken from Ukrainian Wikipedia (https://uk.wikipedia.org/wiki/).

A quantitative report in terms of the number of sentences, text tokens, and types distributed over the different genres is given in Table 1.

Table 1. Text Genre Distribution and Quantitative Data of the UKRMED Medical Text Corpus.

Text Category	Number of sentences	Number of tokens
Complex texts	26,730	329,837
Simple texts	25,395	320,209
Moderate texts	27,081	363,539
Total	79,209	1,013,585

This corpus was created for the experiments with medical text simplification that was launched in [1], and for experiments with readability metrics [11, 12]. We distinguished some featured indices, which are calculated, for our text corpus. The defined features are shown in Table 2. Table 3 shows the corresponding values for the features from Table 2 except "Percentage of speech parts". The analysis of the parts of the speech will be considered separately.

Table 2. Statistic text features.

Text Level	Feature	Abbreviation
Macro level (paragraph level)	Text length in tokens	TLT
	Text length in letters	TLL
Syntactic level	Average sentence length in tokens	ASLT
	Average sentence length in letters	ASLL
Lexical level	Average token length in symbols	ATLS
	Percentage of unique words	US
	Percentage of monosyllables	M
	Average word repetition rate	AWR
	Percentage of speech parts (nouns, verbs, adjectives etc.)	PSP

As we find out such indices as TLT, TLL, ASLT, ATLS, and US from Table 3 have very close values. Small differences have ASLL and M features. These values are not obvious due to the average sentence length is bigger for *"Moderate texts"* or the number of monosyllables (for example, *"жар"/fever*, *"корь"/measles*) is bigger for texts from *"Complex texts"*.

Table 3. Values of statistic text features.

Text Feature	*Complex texts*	*Simple texts*	*Moderate texts*
TLT	329,837	320,209	363,539
TLL	2,051,544	1,925,313	2,313,698
ASLT	7,63	7,23	7,67
ASLL	76,75	75,81	85,44
ATLS	5,87	5,65	6,03
US	4%	5%	4%
M	40%	38%	35%
AWR	11,34	10,35	11,62

The main purpose of the future work with UKRMED is an analysis of readability and text simplification in the medical domain. Caused by our plans we take values of average word length in symbols (ATLS) for all the categories (Fig. 2) and try to analyze these feature values. Fig. 2 shows that the *"Simple texts"* category has the smallest length of words on average but *"Moderate texts"* and *"Complex texts"* have similar values for token length. It helps us to future research for text simplification and assessment of text difficulty.

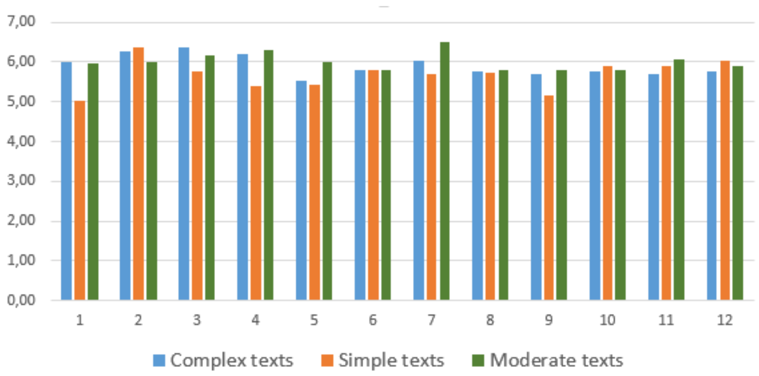

Fig. 2. ATLS values for all categories.

3.3 Part-of-Speech Tagset

We used POS tagger Pymorph2 for the analysis of speech parts of our corpus (https://pymorphy2.readthedocs.io/en/latest/index.html). The part of tagset for the POS annotation of UKRMED Medical Text Corpus is presented in Table 4.

Table 4. The tagset of POS tagger Pymorph2.

POS Tag	Value
NOUN	noun
ADJF	adjective name (full)
COMP	comparative
VERB	verb (personal form)
GRND	gerund
NUMR	numeral
ADVB	adverb
NPRO	pronoun
PRED	predicative
PREP	preposition
CONJ	conjunction
INTJ	interjection
PRCL	particle

As a result of our experiments, we received parts of speech categorization for our three categories (Table 5). Table 5 shows that the most numerous parts are NOUN, ADJF, and VERB. Other parts of speech are rather small in comparison with them.

Table 5. Parts of speech assignment.

POS Tag	*Complex texts*	*Simple texts*	*Moderate texts*	Total
NOUN	130,467	119,641	148,749	398,857
ADJF	57,300	42,173	61,107	160,580
VERB	21,973	36,063	31,013	89,049
INTJ	15	24	11	50
NPRO	7,854	15,762	13,607	37,223
ADVB	7,431	13,371	11,550	32,352
None	3,605	7,091	2,637	13,333
PRCL	805	2,124	1,529	4,458
CONJ	928	1,426	1,839	4,193
PREP	1,079	826	1,416	3,321
PRED	451	910	561	1,922
GRND	466	715	595	1,776
NUMR	426	550	643	1,619
COMP	1,611	-	-	1,611

We have found out also big enough category as *None*. It is a category for words that the morph analyzer could not analyze.

In our work, we tried to analyze the most common words for each category. For this, on the pre-processing stage, all tokens were lemmatized using Pymorphy2. The results are presented in Table 6.

Table 6. Statistic about lemmas for each category
(Lemma – POS – Frequency – Relative frequency).

Complex texts	*Simple texts*	*Moderate texts*
лікування/treatment - NOUN - 1977 - 0.6	який/which/who - NPRO - 2629 - 0.82	який/which/who - NPRO - 2826 - 0.78
який/which/who - NPRO - 1658 - 0.5	цей/this- NPRO - 1551 - 0.48	цей/this - NPRO - 1374 - 0.38
пацієнт/patient - NOUN - 1453 - 0.44	захворювання/disease - NOUN - 1332 - 0.42	може/maybe - None - 1338 - 0.37
хворий/sick - ADJF - 1301 - 0.39	може/maybe - None - 1198 - 0.37	лікування/treatment - NOUN - 1256 - 0.35
дослідження/research - NOUN - 1002 - 0.3	такий/such - NPRO - 1160 - 0.36	також/also - CONJ - 1240 - 0.34
захворювання/disease - NOUN - 925 - 0.28	лікування/treatment - NOUN - 1108 - 0.35	інший/other - NPRO - 1198 - 0.33
цей/this - NPRO - 907 - 0.27	шкіра/skin - NOUN - 1068 - 0.33	захворювання/disease - NOUN - 1162 - 0.32
після/after - ADVB - 865 - 0.26	весь/all - NPRO - 971 - 0.3	такий/such - NPRO - 1106 - 0.3
мати/have - NOUN - 849 - 0.26	вони/they - NPRO - 942 - 0.29	мати/have - NOUN - 1017 - 0.28
даний/given - ADJF - 754 - 0.23	про/about - NOUN - 874 - 0.27	випадок/happening - NOUN - 982 - 0.27

The table shows the top 10 frequency lemmas for each category. From Table 6 it is seen that some of the words are repeated and some are not. Words that are repeated in all categories are highlighted in blue, which are found in *“Complex texts”* and *“Moderate texts”* – in red and in *“Simple texts”* and *“Moderate texts”* – in green.

The category *“Complex texts”* is perfectly characterized by such words as *“treatment”*, *“patient”*, *“patient”*, *“study”*. They are at the top of the list of frequency tokens. Such a word as *“which”* is a frequency word for all three categories, since, in our opinion, sentences in the medical field are quite long and you have to use pronouns and subordinate clauses. Relative pronouns (for example, *“який”/”which/who”*) play the role of conjoined words for joining subordinate clauses to main ones.

The word *”може”/”can”*, which is found on *“Moderate texts”* and *“Simple texts”*, is due to the fact that the discussion of diseases, treatment or diagnosis is advisory. It is also a pronoun and it allows you to highlight objects in a speech situation. Thus, these pronouns connect the text, help to avoid repetition. The word *“мати”* (translated as a noun *"mother"* or a verb *"have"*) was mislabeled. In our case, it is a verb, not a noun. Therefore we have also a disambiguation problem.

4 Conclusion and Future Work

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.

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.

In our paper, we described the structure and the quantitative features of UKRMED, an annotated Ukrainian text corpus that contains three classes of medical texts. On a wide scale, such kind of language resources is a valuable asset for up-to-days researches and the development of effective medical domain targeted technologies. The future study will focus on medical fact extraction from official documents, Ukrainian text simplification in the medical field, question answering services development, etc.

References

1. Cherednichenko, O., Kanishcheva, O., Babkova, N.: Complex term identification for Ukrainian medical texts. In: Proc. of the 1st International Workshop on Informatics & Data-Driven Medicine (IDDM 2018), Vol. 2255, pp. 146–154, CEUR-WS (2018).
2. Gruzitis, N., Pretkalnina, L., Saulite, B., Rituma, L., Nespore-Berzkalne, G., Znotins, A., Paikens, P.: Creation of a balanced state-of-the-art multilayer corpus for NLU. In: Proc. Of the 11th International Conference on Language Resources and Evaluation, pp. 4506–4513 (2019).
3. Janssen, M.: TEITOK: Text-faithful annotated corpora. In: Proc. of the 10th International Conference on Language Resources and Evaluation, LREC 2016, pp. 4037–4043 (2016).
4. Belvin, R. S., May, W., Narayanan, S., Georgiou, P., Ganjavi, S.: Creation of a doctor-patient dialogue corpus using standardized patients. In: Proc. of the 4th International Conference on Language Resources and Evaluation, LREC 2004, pp. 187–190 (2004).

5. Goeuriot, L., Na, J. C., Kyaing, W. Y. M., Khoo, C., Chang, Y. K., Theng, Y. L., Kim, J. J.: Sentiment lexicons for health-related opinion mining. In: Proc. of the 2nd ACM SIGHIT International Health Informatics Symposium, pp. 219–225 (2012).
6. Bott, S., Saggion, H.: An Unsupervised Alignment Algorithm for Text Simplification Corpus Construction. In: Proc. of the Workshop on Monolingual Text-To-Text Generation, pp. 20–26 (2011).
7. Packham, S., Suleman, H.: Crowdsourcing a text corpus is not a game. In: Lecture Notes in Computer Science, Vol. 9469, pp. 225–234 (2015).
8. Cano, C., Monaghan, T., Blanco, A., Wall, D. P., Peshkin, L.: Collaborative text-annotation resource for disease-centered relation extraction from biomedical text. Journal of Biomedical Informatics, 42(5), pp. 967–977 (2009).
9. Velychenko, O., Popova, O.: Cross-cultural specificities of rendering texts on medical ethics in Ukrainian translation. Naukovy Visnyk of South Ukrainian National Pedagogical University Named after K. D. Ushynsky: Linguistic Sciences, 2019(29), pp. 36–50 (2019).
10. Zubatiuk, O., Nosova, E.: Neuroscience-based nomenclature in Ukraine. European Neuropsychopharmacology, 27, S663 (2017).
11. Wermter, J., Hahn, U.: An Annotated German-Language Medical Text Corpus as Language Resource. In: Proc. of the Fourth International Conference on Language Resources and Evaluation (LREC'04), pp. 473-476 (2004)
12. Readability formulas (title from screen), https://readable.com/features/readability-formulas/ last accessed 2020/04/12.
13. Vysotska, V., Lytvyn, V., Burov, Y., Gozhyj, A., Makara, S.: The consolidated information web-resource about pharmacy networks in city. In: Proc. of the 1st International Workshop on Informatics & Data-Driven Medicine (IDDM 2018), CEUR Workshop Proceedings, pp. 239-255 (2018).
14. Lytvyn V., Burov Y., Kravets P., Vysotska V., Demchuk A., Berko A., Ryshkovets Y., Shcherbak S., Naum O.: Methods and Models of Intellectual Processing of Texts for Building Ontologies of Software for Medical Terms Identification in Content Classification. In: CEUR Workshop Proceedings, of the 2nd International Workshop on Informatics & Data-Driven Medicine (IDDM 2019), Vol. 2362, pp. 354-368 (2019).