DOI: 10.15439/2020F192
ISSN 2300-5963 ACSIS, Vol. 21

Automatic Generation of Annotated Corpora of Diagnoses with ICD-10 codes based on Open Data and Linked Open Data

Svetla Boytcheva
Institute of Information and
Communication Technologies
Bulgarian Academy of Sciences, Sofia, Bulgaria
Email: svetla.boytcheva@gmail.com

Boris Velichkov, Gerasim Velchev, Ivan Koychev
Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski", Sofia, Bulgaria
Email: {bobby.velichkov, gerasim.petrov.velchev}@gmail.com
ivan.koychev@fmi.uni-sofia.bg

Abstract—**We propose methods for automatic generation of corpora that contains descriptions of diagnoses in Bulgarian and their associated codes in ICD-10-CM (International Classification of Diseases, 10th revision, Clinical Modification). The proposed approach is based on the available open data and Linked Open Data and can be easily adapted for other languages. The resulted corpora generated for the Bulgarian clinical texts consists of about 370,000 pairs of diagnoses and corresponding ICD-10 codes and is beyond the usual size that can be generated manually, moreover it was created from scratch and for a relatively short time. Further updates of the corpora are also possible whenever new open resources are available or the current ones are updated.**

I. Introduction

THE AUTOMATIC processing and extraction of knowledge from medical texts is a task of public importance. The majority of healthcare documents are still available mainly in free text format, on the local language. Natural Language processing of clinical text require to be developed specific language resources that requires expert knowledge and validation, which is quite difficult to achieve, especially in a situation where health workers are overwhelmed with other more important daily responsibilities. Clinical Natural Language Processing (NLP) is quite challenging task for non-English language [1]. For low resource languages such as Bulgarian the majority of the required resources are not available, or there are some limited versions. The question is "how we can develop automatically or semi-automatically such resources from scratch for relatively short time?". Medical terminology in Bulgarian has very specific nature, because it is a mixture between terminology in Bulgarian, Latin and transliterated Latin terms in Cyrillic [2].

Diagnosis is one of the most important complex data on the patient's health in clinical texts. On the other hand, due to the complexity of the information they contain, there is a wide variety of ways to describe it - using different terms, paraphrases, abbreviations and various details to describe the stage of the disorder, its location, cause and severity,

Further processing of the extracted diagnosis information requires unification/normalization of the data according to some standard nomenclatures to avoid ambiguities. One of the widely used International Classification of Diseases is ICD-10 [1] that has also translations to many languages. The classification task for association of ICD-10 codes to textual descriptions of diagnosis requires training corpora, and because there are about 11,000 different codes the corpora should be relatively large in size. We focused on the task for automatic generation of training corpora of diagnosis descriptions in Bulgarian and their corresponding ICD-10 codes.

Already there are various research for other languages. Wang, Qiong, et al. [5] present a study that aims to develop and evaluate effective methods that can normalize diagnosis and procedure terms written by physicians to standard concepts in ICD in Chinese using an entity-linking framework and two manually annotated datasets (8,547 diagnoses and 8,282 procedures). Marovac, Avdić et al. [6] present in their paper the process of creating medical lexical resources for the Serbian language and they achieve mapping to certain ICD-10 codes with precision over 80%. Almagro, Unanue et al. [7] have been carried out an exploration on 7254 Spanish hospital discharge reports for a period of 3 years resulting in total 76,525 identified codes with approximately 7,000 unique ones. Bagheri, Sammani et. al. [8] sought to implement a system to help 3-digit Dutch ICD-10 coding of discharge letters via machine learning algorithms. Dalianis [9] addresses the automatic assignment of Portuguese ICD-10 codes for causes of death by analyzing 114,228 free-text descriptions containing a total of 1,418 distinct codes.

We propose methods for automatic generation of corpora that contains descriptions of diagnoses in Bulgarian and their associated codes in ICD-10. The proposed approach is based on the publicly available resources and can be easily adapted

This research is partially funded by the Bulgarian Ministry of Education and Science, grant DO1-200/2018 'Electronic health care in Bulgaria' (e-Zdrave) and the Bulgarian National Science Fund, grant DN-02/4-2016 'Specialized Data Mining Methods Based on Semantic Attributes' (IZIDA). We are grateful to anonymous reviewers for useful comments and suggestions.

[1] https://icd.who.int/browse10/2019/en

for other languages. The resulted corpora generated for the Bulgarian clinical texts consists of about 370,000 pairs of diagnosis and the corresponding ICD-10 codes and is beyond the usual size that can be generated manually, moreover it was created from scratch and for a relatively short time. Further updates of the corpora are also possible whenever new open resources are available or the current ones are updated.

II. Method

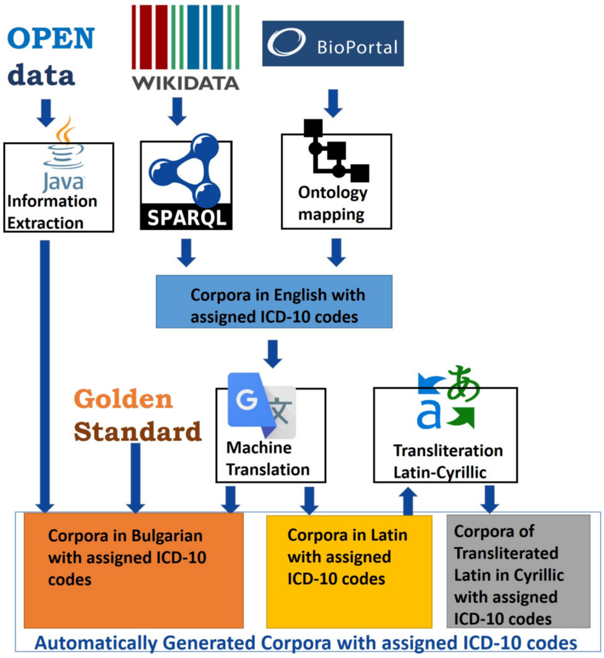

OPEN data
WIKIDATA
BioPortal
Java
Information Extraction
SPARQL
Ontology mapping
Corpora in English with assigned ICD-10 codes
Golden Standard
Machine Translation
Transliteration Latin-Cyrillic
Corpora in Bulgarian with assigned ICD-10 codes
Corpora in Latin with assigned ICD-10 codes
Corpora of Transliterated Latin in Cyrillic with assigned ICD-10 codes
Automatically Generated Corpora with assigned ICD-10 codes

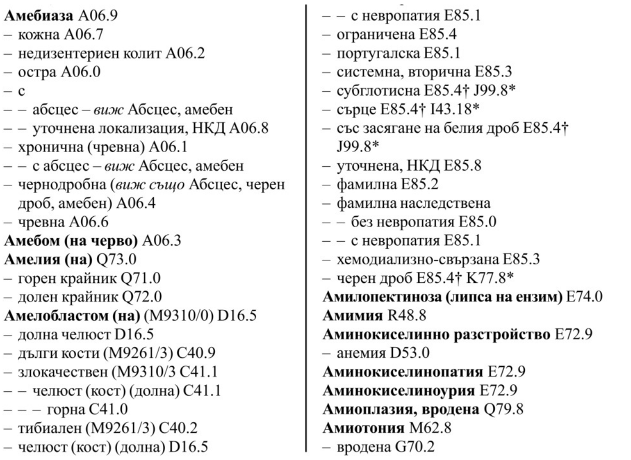
Figure 1. Method

The proposed method for automatic corpora generation is language independent and relies mainly on open data and linked open data (LOD[2]). The following components are used:

- Automatic extraction of Annotations from Open Documents - for this module are used publicly available documents in Bulgarian language as an input, and are developed information extraction algorithms that convert the textual data into dataset of structured pairs of diagnosis and associated ICD-10 codes;
- Automatic extraction of Annotations from LOD - for this module are used SPARQL queries for extraction of diagnosis in English language and corresponding codes to some of the widely used standard classifications. All available mappings between these classifications are used to produce dataset with associated codes to ICD-10.
- Machine translation - this module is used for diagnosis translation from English to Bulgarian and Latin.
- Transliteration tool - this module is responsible for transliteration of the diagnosis from Latin to Cyrillic.
- Other resources - Golden Standard (GS) for some diagnoses with associated ICD-10 codes [3].

A. *Automatic Extraction of Annotations from Open Documents*

The main resource used in this module is the official document of the International Classification of Diseases: "ICD-10-CM Alphabetical Index" (ICD10-Index[3]). This document is translated by health organizations or ministries into the relevant language. A translated version[4] from the website of the Ministry of Health of the Republic of Bulgaria was used to generate the current dataset. It is in the form of two PDF documents. First they are converted to Microsoft Word (".doc") format using the PDF reader "Nitro Pro (7.5.0.18)"[5]. The resulting format is converted in addition to the ".docx" format using Microsoft Word functions. The Apache POI[6] library with the Java programming language are used to read the received documents. The library provides an easy way to read the individual paragraphs. It is important to note that in the process of conversion some paragraphs can be damaged, thus some minor manual cleaning and formatting of the result file is needed. For the rest of the (automatic) part of the processing[7], the structure of the index is very important for the rule based information extraction. A screenshot of the document can be seen at Figure 2.


Амебиаза A06.9
– кожна A06.7
– недизентериен колит A06.2
– остра A06.0
– с
– – абсцес – *виж* Абсцес, амебен
– – уточнена локализация, НКД A06.8
– хронична (чревна) A06.1
– – с абсцес – *виж* Абсцес, амебен
– чернодробна (*виж също* Абсцес, черен дроб, амебен) A06.4
– чревна A06.6
Амебом (на черво) A06.3
Амелия (на) Q73.0
– горен крайник Q71.0
– долен крайник Q72.0
Амелобластом (на) (M9310/0) D16.5
– долна челюст D16.5
– дълги кости (M9261/3) C40.9
– злокачествен (M9310/3 C41.1
– – челюст (кост) (долна) C41.1
– – – горна C41.0
– тибиален (M9261/3) C40.2
– челюст (кост) (долна) D16.5

– – с невропатия E85.1
– ограничена E85.4
– португалска E85.1
– системна, вторична E85.3
– субглотисна E85.4† J99.8*
– сърце E85.4† I43.18*
– със засягане на белия дроб E85.4† J99.8*
– уточнена, НКД E85.8
– фамилна E85.2
– фамилна наследствена
– – без невропатия E85.0
– – с невропатия E85.1
– хемодиализно-свързана E85.3
– черен дроб E85.4† K77.8*
Амилопектиноза (липса на ензим) E74.0
Амимия R48.8
Аминокиселинно разстройство E72.9
– анемия D53.0
Аминокиселинопатия E72.9
Аминокиселиноурия E72.9
Амиоплазия, вродена Q79.8
Амиотония M62.8
– вродена G70.2

Figure 2. ICD-10 Alphabetical Index

To avoid unnecessary repetition, the index is organized in the form of a tree structure: leading terms, which are located in the leftmost column, and other paragraphs, which start on the right. For this reason, the full term consists of several lines, sometimes giving too broad description.

When traversing each node in such tree structure, a number of regular expressions are used in order to be able to determine both the level of the text in the tree and to recognize the individual text items. Some of the main text processing transformations are the following:

- Convert all references to pre-specified categories: "виж също -> виж" (see also -> see).

[2] https://lod-cloud.net/

[3] https://icd.codes/icd10cm/alphabetical-index

[4] https://ncpha.government.bg/bg/2019-02-19-23-22-18/icd-10

[5] https://www.gonitro.com/nps/pro/pdf-software

[6] https://poi.apache.org/

[7] https://github.com/BorisVelichkov/ICD10-Medical-Data

- Remove noun inflexion forms as number and case: "(-а)", "(-та)", etc.
- Remove parentheses, other special characters and redundant white spaces.
- Merge words that have been transferred to a new line.
- Merge erroneously separated paragraphs.
- Combine the different levels in document structure order to form correct sentences for diagnosis.
- Recognize references and remove them after concatenation with the next level text.
- Recognize ICD-10 codes and create valid examples for each type of ICD-10 code (the codes are written in a different format).

Some examples of the diagnosis descriptions with the corresponding ICD-10 codes generated from the tree structure (Figure 2) are displayed in Table I.

Table I
ICD-10 INSTANCES CREATED FROM THE SHOWN TREE STRUCTURE

ICD10	Text
A06.9	Амебиаза
A06.7	Амебиаза кожна
A06.2	Амебиаза недизентериен колит
A06.0	Амебиаза остра
A06.8	Амебиаза с уточнена локализация
A06.1	Амебиаза хронична чревна
A06.4	Амебиаза чернодробна виж Абсцес черен дроб амебен
A06.6	Амебиаза чревна

B. Automatic Extraction of Annotations from LOD

The main resource are translation of the ICD-10-CM translated in Bulgarian[8] (ICD10-BG). It contains (see Fig. 1) about 11,000 classes organized in 4 levels hierarchy - 22 groups at level 1, 211 subgroups at level 2, 2025 are level 3 (3-sign codes) and 8946 at level 4 (4-sign codes). They are not presenting single diagnose, but statistical classification of groups of diagnoses. Thus they can serve only partially as a resource for the generated corpora. The ICD-10 is one of the widely used classification of diseases and translations[9] on several languages are available.

Table II
WIKIDATA ONTOLOGIES

Wikidata code	Q12136	Q179630	Q169872	Q639907
P4229	39,743	107	20	10
P699	47,092	60	36	13
P486	27,478	127	207	39
P3841	5,952	2	100	68
P604	6,326	37	85	20
P5270	50,292	101	55	37
P1550	31,179	103	15	13

As primary resource was used Wikidata[10], which provides encyclopedic data in structured format. Unfortunately only for small subsets of diagnosis are available labels in Bulgarian language, thus our primary focus will be to collect data for English language. We collect from Wikidata results of several SPARQL queries investigating for labels in English language, the availability of the concepts disease (Q12136), illness (Q814207), syndrome (Q179630), symptom (Q169872), medical finding (Q639907) and their associations with medical classifications: ICD-10-CM (P4229), Human Disease Ontology[11] (P699), MeSH[12] (P486), The Human Phenotype Ontology[13](P3841), MedlinePlus (P604), MonDO[14] (P5270), Orphanet[15] (P1550).

The results from different combinations of concepts and ontologies are shown in Table (Table II), where also mapping of the ontologies ID to ICD-10 code is applied (if any). The SPARQL queries[16] were run in Wikidata Query Service and the generated results are stored in CSV format including the following properties *<Item URI to Wikidata, Item Label, Item Alternative Label, Ontology ID, ICD10 code>*. All result CSV tables are merged, and are removed duplicates. For some labels, that contain disjunction ("or") a separate instance is create for each element. The total collected datasets contains 57,142 pairs of 4-sign code in ICD-10 and text label for diagnosis. Further automatic cleaning was applied to remove abbreviations. For example, "ID" caused in normalization some ambiguities and was misinterpreted as "Identification document", instead of "Infectious disease". The final result cleaned dataset (WD-ENG) contain 55,292 pairs of data with ICD-10 codes (4-sign) and diagnosis in English Language.

C. Latin-Cyrillic Transliteration

One of the most important parts of building a usable dataset is data augmentation - the process of generating new data as a variation of already known. In medical text such variations include different ways of writing diagnosis names: using Bulgarian terms, Latin terms, or using Latin terms written using Cyrillic letters. Such variants we call Cyrillic transliterations of Latin terms. There exist a plenty of rules for transliteration of Latin medical terms in Cyrillic representations, described in a Latin-Bulgarian dictionary [4]. We categorise them in three groups depending on number of consecutive Latin letters (1, 2 or 3) they are replacing with string in Cyrillic (type 1, type 2, or type 3 respective). There are 22 rules of type 1, 11 rules of type 2 and 9 from of type 3. Some of the rules are direct replacements of a string with another string. Other include wildcard positions - positions which could be replaced with a set of symbols (e.g. vowels). For instance, rules of:

- type 1 - "u" ⇒ "у", "x" ⇒ "кс".

[8] http://www.zdrave.bg/normativi/MKB10.pdf

[9] https://www.who.int/classifications/icd/ICD-10\%20languages.pdf

[10] https://www.wikidata.org

[11] https://www.ebi.ac.uk/ols/ontologies/doid

[12] https://www.nlm.nih.gov/mesh/meshhome.html

[13] https://hpo.jax.org/app/

[14] https://mondo.monarchinitiative.org/

[15] http://www.orphadata.org/cgi-bin/index.php

[16] https://w.wiki/ZZo

- type 2 - "ci" ⇒ "ци", "ch" ⇒ "х", "qu" ⇒ "кв".
- type 3 - "sua" ⇒ "сва", "sui" ⇒ "суи", "sm" + vowel ⇒ "зм", "rs" + vowel ⇒ "рз"

According to the rules, "basis" transliterates to "базис", "trapez" to "трапец", "xantos" to "ксантос", "sensibilis" to "сензибилис", "neoplasma" to "неоплазма", "suillus" to "суилус", "xiphos" to "ксифос", etc.

The algorithm of transliteration[17] consists of parsing the input string, according to the rules, recognizing groups from the left parts of the rules and generating the output string replacing the left parts with their corresponding right parts. The order in which we applied the rules is from the ones with longest context to the ones with shortest context because the longest ones are more specific and some of the shortest could be their subset, so we give priority to the specificity.

Some of the rules depend on the origin of the word, Greek or Latin, and they are applied only to words with specific origin[18]. For instance, rule only for words with Greek origin is "ph" ⇒ "ф", and rule only for words in Latin origin is "z" ⇒ "ц". So we should create a naive origin extractor. For this purpose we use a corpus of typical prefixes, suffixes and roots of words with a Greek origin and words with a Latin origin . Typical prefixes of words with Greek origin are, e.g., "rhe-", "xanth-", "zon-" and typical prefixes of words with Latin origin are "sub-", "form-", "celer-". In order to extract the origin of a word, we should count the number of prefixes, number of suffixes and number of roots which it contains, respectively for the Greek and Latin ones, and then define the origin to be the one which is more often contained.

A plenty of diseases contain in their notation the name of their founder. For the sake of simplicity, we apply the same rules to the transliteration of names. This is a downside because they should be transliterated according to the transliteration rules of the language they originate from.

Letter "w" does not exist in Latin. It could be transliterated in either German ("в") or English ("у") style. E.g., "Kwashiorkor" should be transliterated to "Квашиоркор", but "Williams" to "Уилиямс". In order not to make the algorithm more complicated using origin extractor of names and applying different rules to name transliteration, we use the German style by default because it occurs more often.

Numbers in Roman notation (they consist of Latin letters) should not be transliterated.

III. Corpora Generation

Further was applied Machine translation using Google Translation of the data from English→Bulgarian (WD-BG); English→Latin (WD-LAT); and transliteration of the result data set in Latin to Cyrillic (WD-TRANS) applying methods described in the next section.

As a result of the creation of datasets, 6 datasets have been generated (see Figure III. There are two options for each of them: one with 4-sign codes and one with 3-sign codes. All have format: <*(ICD-10, Text*>. In the process of merging each of the datasets, the following main transformations are applied:

- Transformation of all homoglyphs.
- Remove and convert all obviously incorrectly written codes so that they become valid codes (for example, unnecessary blanks are removed).
- Remove all duplicates.
- Removal of all codes that are not in compliance with the official list of codes used in Bulgaria to date (codes from ICD10-BG).

Table III
ICD10 3 sign and 4 sign datasets statistics

Dataset	**Total Instances**		**Unique Codes**	
	3-sign	**4-sign**	**3-sign**	**4-sign**
ICD10-BG	2025	8946	2025	8946
GS	409	4212	42	408
ICD10-Index	2176	42811	310	8420
WD-BG	3879	46686	434	3499
WD-LAT	3879	46686	434	3499
WD-TRANS	3879	46686	434	3499
Corpus	**189756**	**383042**	**2035**	**10971**

More detailed statistics can be seen in the Table III and the contribution of each dataset to the generated corpus[19] (see Figure III) and the distribution of instances per classes (see Figure II-C). It is important to mention that “Corpus-4Sign” contains all examples from “Corpus-3Sign”, because all valid 3 sign codes are valid 4 sign codes too.

We can state that the resulting corpus is valid because only official or trusted sources were used to create it and there are no personally associated codes. The open data (ICD10-Index) is an official document from the website of the Ministry of Health in Bulgaria. The Linked Open Data (ICD10-BG) is also 100% reliable as it represents official classifications and ontologies. The gold standard (GS) is made by doctors, which in itself ensures that it is a reliable source. The data taken from Wikidata (WD-LAT and WD-BG) are not an official document, but we can say that they are trusted, as they have been checked through several types of classifications before being approved for publication online. The data generated by transliteration (WD-TRANS) are valid because the rules described in the necessary literature for this are strictly applied for their generation.

IV. Conclusion and Further Work

The paper presents a method for automatic generation of corpora that contains descriptions of diagnoses in Bulgarian and their associated codes in ICD-10-CM. The proposed approach is based on the available open data and Linked Open Data. The proposed approach employs methods for

[17] https://github.com/BorisVelichkov/latin-transliterator

[18] https://www.oakton.edu/user/3/gherrera/Greek\%20and\%20Latin\%20Roots\%20in\%20English/greek_and_latin_roots.pdf

[19] https://github.com/BorisVelichkov/ICD10-Medical-Data/tree/master/datasets

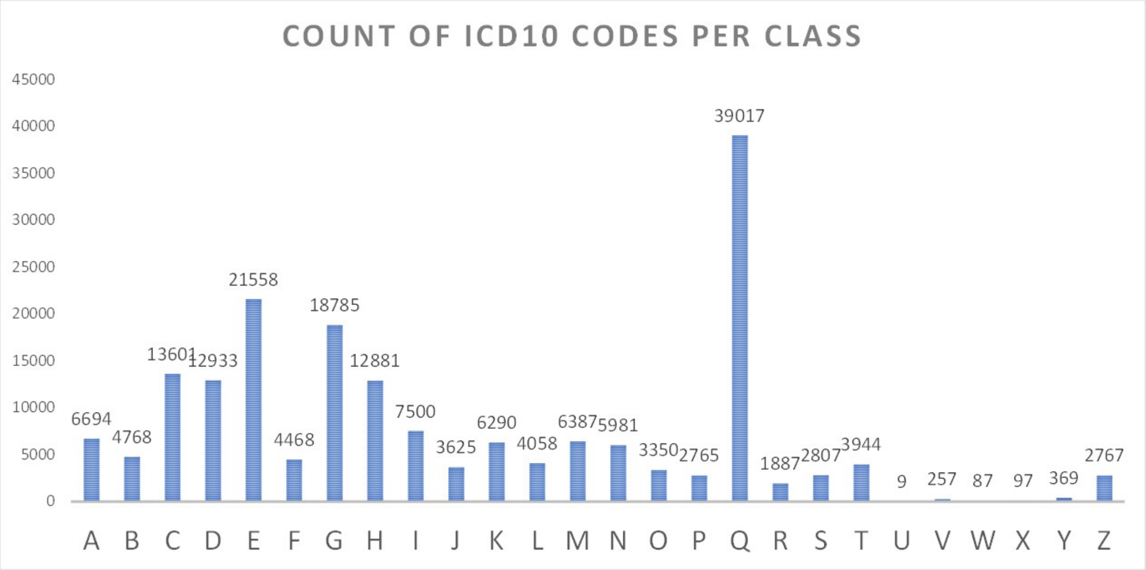


Figure 3. Diagnosis descriptions per ICD-10 code class in the generated corpora

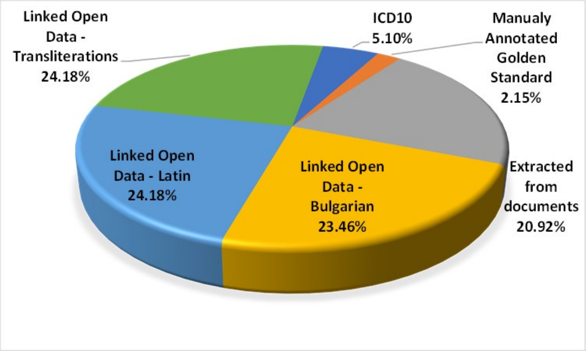


Figure 4. Contribution of each resources to the generated corpora

automatic terms and relations extraction from semi-structured documents; methods for automatic terms extraction from Linked Open Data Cloud and suggests techniques for Latin-Cyrillic transliteration. The resulted corpora generated for the Bulgarian clinical texts consists of about 370,000 pairs and is beyond the usual size that can be generated manually, moreover it was created from scratch and for a relatively short time. Up to our knowledge this is the largest dataset of this type. Further updates of the corpora are also possible whenever new open resources are available or the current ones are updated. The proposed approach is relatively language independent and can be easily adapted for other languages.

Since the generated corpus is highly unbalanced, it is good to do a Data Augmentation [10] in order to reduce the more drastic differences in the number of individual classes. One possible option that would be applicable in the current dataset is through the use of synonyms [11].

REFERENCES

[1] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, P. Zweigenbaum. "Clinical natural language processing in languages other than english: opportunities and challenges." Journal of biomedical semantics, 2018 Dec 1;9(1):12.

[2] S. Boytcheva, "Multilingual aspects of information extraction from medical texts in Bulgarian." Multilingual Processing in Eastern and Southern EU Languages: Less-resourced Technologies and Translation, Cambridge Scholars Publishing. 2012 Apr 25:308-29.

[3] S. Boytcheva, "Automatic matching of ICD-10 codes to diagnoses in discharge letters."*In Proceedings of the second workshop on biomedical natural language processing, RANLP 2011*, pp. 11-18, September 2011.

[4] M. Voinov et al. *Latin-Bulgarian Dictionary*. Planeta-3, pp. 792, 1999. (in Bulgarian)

[5] Q. Wang et al. "A study of entity-linking methods for normalizing Chinese diagnosis and procedure terms to ICD codes". Journal of Biomedical Informatics. 2020 Apr 13:103418. https://doi.org/10.1016/j.jbi.2020.103418

[6] U. Marovac, A. Avdić, D. Janković, and S. Marovac. "Creating Resources for Marking Diagnoses in Electronic Health Reports in Serbian". International Journal of Electrical Engineering and Computing, 2020. 4(1), pp. 18-23.

[7] M. Almagro, R. M. Unanue, V. Fresno and S. Montalvo, "ICD-10 Coding of Spanish Electronic Discharge Summaries: An Extreme Classification Problem", IEEE Access, 2020, vol. 8, pp. 100073-100083, 2020, doi: 10.1109/ACCESS.2020.2997241.

[8] A. Bagheri, A. Sammani, PGM Van der Heijden, FW Asselbergs, and DL Oberski. "Automatic ICD-10 classification of diseases from Dutch discharge letters". In: Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 3: C2C. 2020, pp. 281-289.

[9] H. Dalianis. "Clinical text retrieval-an overview of basic building blocks and applications". In Professional Search in the Modern World, 2014, pp. 147-165. Springer, Cham.

[10] J. Wei, and K. Zou. "Eda: Easy data augmentation techniques for boosting performance on text classification tasks". arXiv preprint arXiv:1901.11196. 2019 Jan 31.

[11] N. Khairova, S. Petrasova, W. Lewoniewski, O. Mamyrbayev, and K. Mukhsina. "Automatic extraction of synonymous collocation pairs from a text corpus". In 2018 Federated Conference on Computer Science and Information Systems (FedCSIS)". 2018 Sep 9, pp. 485-488, IEEE.