Automatic Transformation of Clinical Narratives into Structured Format

Sylvia Vassileva[1], **Gergana Todorova**[1], **Kristina Ivanova**[1], **Boris Velichkov**[1,2], **Ivan Koychev**[1,2], **Galia Angelova**[3] and **Svetla Boytcheva**[3]
[1] Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", Sofia, Bulgaria
sylvia.vassileva@gmail.com, gergana.d.todorova@gmail.com,
kiivanova735@gmail.com, koychev@fmi.uni-sofia.bg
[2] GATE Institute, Sofia University "St. Kliment Ohridski", Sofia, Bulgaria
boris.velichkov@gate-ai.eu
[3] Institute of Information and Communication Technologies,
Bulgarian Academy of Sciences, Sofia, Bulgaria
galia@lml.bas.bg, svetla.boytcheva@iict.bas.bg

Abstract

Vast amounts of data in healthcare are available in unstructured text format, usually in the local language of the countries. These documents contain valuable information. Secondary use of clinical narratives and information extraction of key facts and relations from them about the patient disease history can foster preventive medicine and improve healthcare. In this paper, we propose a hybrid method for the automatic transformation of clinical text into a structured format. The documents are automatically sectioned into the following parts: diagnosis, patient history, patient status, lab results. For the "Diagnosis" section a deep learning text-based encoding into ICD-10 codes is applied using MBG-ClinicalBERT - a fine-tuned ClinicalBERT model for Bulgarian medical text. From the "Patient History" section, we identify patient symptoms using a rule-based approach enhanced with similarity search based on MBG-ClinicalBERT word embeddings. We also identify symptom relations like negation. For the "Patient Status" description, binary classification is used to determine the status of each anatomic organ. In this paper, we demonstrate different methods for adapting NLP tools for English and other languages to a low resource language like Bulgarian.

1 Introduction

The analysis of medical data can lead to the discovery of knowledge that has a significant effect on healthcare and the effectiveness of treatment. Unfortunately, large amounts of health data are available in an unstructured text format, usually in the local language of the countries. These documents contain valuable information, but it cannot be extracted directly from them. Transformation of the clinical record from unstructured text format into a structured presentation that includes key facts and relations will enable further analysis, which can help to obtain valuable information about the patient's medical history and in this way to support preventive medicine and improve healthcare.

Vaci et al. (2020) use huge manually annotated corpora for patients with depression with expert-selected terms of interest. They apply statistical models and active learning for structured information extraction with an average F1-score of 74% for temporal characteristics, diagnosis, disease history, symptoms, cognitive scores, medications, response, and adverse effects. Taira et al. (2001) also apply statistical methods for structuring information from radiology reports. Denecke (2008) presents a method SeReMeD, based on semantic transformation rules that achieves precision 93% and recall 88% for chest X-ray reports. Kreimeyer et al. (2017) investigate in their systematic review the clinical NLP methods. Still, the rule-based approaches predominate.

Zhan et al. (2021) apply different word embedding techniques for automatic prediction of ICD-10 codes for free-text clinical notes for 8 cardio-vascular diseases. The reported AU-ROC rating is in the range from 94% up to 99%. Chen et al. (2021) apply hybrid methods combining keyword-based method and deep learning methods (word2vec and sentence2vec), achieving F1-score 83% and 88% correspondingly for both methods.

There is limited research for languages other

Proceedings of the Student Research Workshop associated with RANLP-2021, pages 219–227,
held online, Sep 1–3, 2021.
https://doi.org/10.26615/issn.2603-2821.2021_030

than English about NER from clinical text Névéol et al. (2018). Lei et al. (2014) present NLP techniques for Name Entity Recognition (NER) for clinical text in Chinese. The authors explore the application of different classical machine learning techniques like conditional random fields (CRF), support vector machines (SVM), maximum entropy (ME), and structural SVM (SSVM) over about 400 manually annotated admission notes and discharge letters. The best score is 93% for the former and 90% for the latter datatype. Zhang et al. (2018) propose an entropy-based model for temporal information extraction for medical facts and medical events from Chinese clinical text.

Some preliminary work was reported for structuring clinical information in Bulgarian - for extraction of various numerical values related to lab test by Boytcheva et al. (2015), medication extraction by Boytcheva (2011), and annotation with ICD-10 codes by Velichkov et al. (2020a). The automatic generation of diabetes register from unstructured clinical texts using rule-based NLP techniques is presented in Tcharaktchiev et al. (2018).

Therefore, we aim to develop a Natural Language Processing (NLP) pipeline for automatically extracting information from clinical texts and presenting it in a structured format. This task is even more challenging for low-resource languages like Bulgarian. The developed approach employs both classical methods such as those based on dictionaries and rules and modern language models based on Deep Neural Networks Transformers such as BERT (Bidirectional Encoder Representations from Transformers by Devlin et al. (2018)) and can be easily adapted to other languages.

2 Data

2.1 Diagnosis Dataset

The ICD-10 classification [1] is a hierarchical classification of diseases that consists of multiple levels. Each diagnosis is a description of a disease and is associated with one or more ICD-10 codes. Categories using 3 and 4 signs are used in the Bulgarian healthcare system to encode the diseases. For example, the 3-sign code "C00" is used to encode "Malignant neoplasm of lip", and its 4-sign child "C00.0" encodes "Malignant neoplasm of external upper lip". The ICD-10 classification has 2,035 3-sign classes and almost 11,000 4-sign classes. In this paper, we classify each diagnosis by choosing the most relevant 4-sign classes.

For the purposes of ICD-10 diagnosis classification task, a dataset created by Velichkov et al. (2020a) was used with additional cleaning and preprocessing to reduce the noise. It contains 345,591 sample diagnoses with their corresponding ICD-10 codes.

2.2 Bulgarian Medical Articles Dataset

We compile a dataset of Bulgarian medical articles starting with a dataset from Velichkov et al. (2020a) and enhancing it by scraping Bulgarian medical websites, containing general medical articles. The final dataset consists of about 10,000 articles. The Bulgarian Medical Articles Dataset is used in the Diagnosis encoding task.

2.3 Patient History Dataset

The Patient History Dataset is used for evaluation purposes and consists of 30 samples of manually annotated patient history records from outpatient records. The records are anonymized and contain 170 sentences, 582 entity token instances across 9 categories. It contains 446 tokens linked to WikiData entities for 111 unique entities. The Patient History Dataset is used in the Patient History information extraction task.

2.4 Patient History Entity Knowledge Base

As part of the Patient History extraction task, we perform entity linking of recognized entities against a knowledge base. We compile a knowledge base from WikiData symptoms, clinical signs, organs, including parts of the body, and anatomical systems which represent UMLS[2] or MeSH[3] ontology concepts. In total, we collected 40,417 entities with 169,359 aliases and translated automatically all aliases in Bulgarian using Google translate. Some data cleaning was performed on the result translation. Unfortunately, there is no official reliable translation of UMLS or MeSH in Bulgarian.

2.5 Patient Status Dataset

For the task of patient status classification, we use a dataset of anonymized patient records. The records include patient status from General Practitioners (GPs) outpatient records. Each sentence in the

[1] https://icd.who.int/browse10/2010/en

[2] https://www.nlm.nih.gov/research/umls/index.html

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

dataset is manually annotated with the organ it refers to and whether or not the organ is healthy or abnormal. The dataset consists of 7,554 sample sentences.

The patient status record contains information about basic patient exams like blood pressure, pulse, Succusio renalis, BMI (Body Mass Index), Height, Waist, Weight, VOD/VOS (Visio oculus dexter/Visio oculus sinistra - visual acuity of the right and left eyes), as well as a description of the patient's symptoms based on a doctor's visual examination. For example, the status may contain "УНГ - зачервено гърло" *("ENT - sore throat")* or *"RR:120/80" (Blood pressure measured using Riva-Rocci method: 120/80).*

3 Methods for Clinical Text Structuring

3.1 Information Extraction from Medical Texts in Bulgarian

This work aims to develop a prototype system for recognizing and extracting medical concepts from clinical texts in Bulgarian. The developed NLP approach consists of the following modules:

- Document sectioning - The discharge letters in Bulgaria follow a standard structure of sections. The documents are automatically divided into the following parts: diagnosis, patient history, patient status, lab test results.

- Diagnosis encoding into ICD-10 - For the "Diagnosis" section, we use MBG-ClinicalBERT model - based on ClinicalBERT, fine-tuned for Bulgarian medical text, to automatically generate ICD-10 codes of diagnoses from the row text.

- Patient History information extraction - From the patient history section, we identify the patient's symptoms using a rule-based approach, improved by similarity search, based on MBG-ClinicalBERT word embeddings. We also identify relationships between symptoms including negation.

- Patient status classification - For the patient status description, binary classification is used to determine the condition of each anatomic organ.

- Lab test results extraction - The laboratory results section is also processed to extract the test and result values for the patient. In this section, we use a rule-based approach.

As smaller contributions, we also consider the presented below various methods for adapting the NLP tools for English and other widely spread languages to a low-resource language such as Bulgarian.

3.2 Automatic Encoding of Diagnosis into ICD-10 codes

Most of the medical diagnoses are presented in discharge letters as unstructured text and concept normalization using a standard classification or ontology needs to be applied for disease identification. The task for automatic encoding of diagnosis into ICD-10 codes can be considered as an extreme multi-class classification task with about 2,000 classes of 3-sign codes and nearly 11,000 classes of 4-sign codes, as well as multi-label, as there can be more than one valid code for a diagnosis.

3.2.1 Fine-Tuning MBG-ClinicalBERT Language Model

For this task we develop MBG-ClinicalBERT - a model based on ClinicalBERT (Alsentzer et al., 2019). ClinicalBERT is a deep learning language model, a variant of BERT (Bidirectional Encoder Representations from Transformers proposed by Devlin et al. (2018)), originally trained on approximately 2 million clinical notes in English. ClinicalBERT is trained on clinical texts only in English and therefore requires additional training to be applied to Bulgarian texts.

The MBG-ClinicalBERT model we develop is based on ClinicalBERT and additionally trained on about 10,000 medical articles from the Bulgarian Medical Articles Dataset (Section 2.2) containing descriptions of diseases in Bulgarian to improve the model's understanding of Bulgarian medical vocabulary and its context use. We split the dataset and use 80% for training, 10% for hyper-parameter tuning, and 10% for testing. The training is performed using the masked language model task by masking 15% of the tokens in the text and training the model to predict them. We use WordPiece tokenization and the original vocabulary of the model and report the resulting perplexity of the model.

3.2.2 Fine-Tuning ICD-10 Diagnosis Classifier

The MBG-ClinicalBERT model is then fine-tuned for the ICD-10 diagnosis classification task by using the standard classification architecture proposed by Devlin et al. (2018). We use the output [CLS] token from MBG-ClinicalBERT, add a softmax linear layer on top, and train it to predict the ICD-10 code using the Diagnosis Dataset from Section 2.1. As a result, we show the top 5 classes with the highest probability. The accuracy, macro-F1 and Mean Reciprocal Rank (MRR) of the model are reported.

The MRR is the mean reciprocal rank and its formula is shown below, where Q is the number of test samples and rank is the rank position of the first relevant document[4]:

$$MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{rank_i} \quad (1)$$

3.3 Hybrid Method for Information Extraction from Patient History

There are different approaches to recognizing objects in text. The main types are dictionary-based, rule-based, using machine learning - supervised or unsupervised methods.

The task of information extraction from the patient history section is performed using a hybrid approach combining a rule-based named entity recognizer and an entity linker working with a knowledge base of symptoms, complaints, organs, and anatomical systems. This module of the system should be able to recognize patient disease history, as well as risk factors, and family history. The system must also recognize connections between symptoms, complaints, and anatomical organs and systems, as well as detect negations of complaints and symptoms.

The overall architecture (Fig. 1) is using a spaCy stanza pipeline [5] [6] which provides sentence splitting, tokenization, POS tagging, lemmatization, and dependency parsing for Bulgarian out of the box. The stanza pipeline is trained using BulTreeBank universal dependencies treebank [7] and is available for direct usage.

3.3.1 Patient History Named Entity Recognition

A custom named entity recognition (NER) component is created using a rule-based approach. The component recognizes entities in 11 categories (Table 1). The idea of the developed method is to find simpler objects through dictionaries and then, with the help of rules, find the connections between the individual objects and identify the concepts sought. Using spaCy's Entity Ruler, the NER matches phrases generated using entity lexicons and rules of several types: matching the lexicon phrases, matching the lexicon phrases combined with adjectives or other parts of speech, matching negation of risk factors, complaints, or symptoms. The rules have been crafted manually by analyzing the Patient History Dataset from Section 2.3. spaCy's matcher automatically matches entities using the predefined rules. The entity category lexicons are compiled manually from the Patient History Dataset (Section 2.3) and WikiData[8] results for the respective categories.

Some of the entity categories can contain nested categories. For example, the family history usually contains symptoms of the family members, and negations can be applied to symptoms, complaints, or risk factors. Therefore, the rules for these categories are generated using previously detected sub-entities. Identifying part of the nested entities is also valuable so it is counted as a correctly recognized token.

For the NER task we report the token-based accuracy of the approach.

3.3.2 Patient History Entity Linking

We use the fine-tuned MBG-ClinicalBERT language model as an encoder to generate word embeddings for each alias in the knowledge base from Section 2.4. The fine-tuning process was described in Section 3.2. The word embeddings are generated using the mean pooling strategy of the last four layers of the BERT model with the help of HuggingFace transformers library[9]. Since BERT provides contextualized word embeddings, the pipeline should be able to use the sentence context information carried in the embeddings to disambiguate entities. To link each recognized entity

[4] https://en.wikipedia.org/wiki/Mean_reciprocal_rank

[5] spaCy https://spacy.io/

[6] spaCy Stanza https://github.com/explosion/spacy-stanza

[7] BulTreeBank https://universaldependencies.org/treebanks/bg_btb/index.html

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

[9] https://huggingface.co/

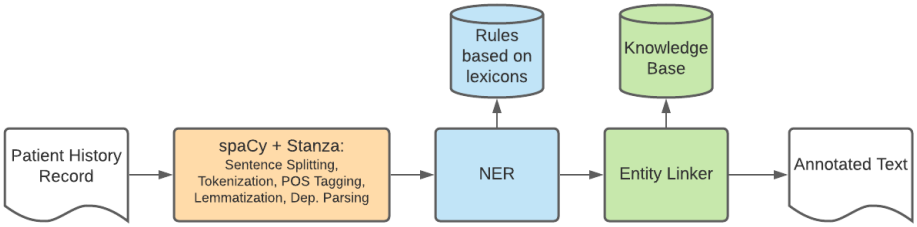

Figure 1: Architecture of the patient history pipeline.

Table 1: Entity categories recognized by the NER component

Category	**Description**	**Lexicon Phrase Count**
ORGAN	Human Organs	1200
COMPLAINT	A patient's subjective report of their problems	1110
SYMPTOM	A doctor's assessment of the patient's problems (supported by clinical signs)	1646
ANATOMICAL SYSTEM	Anatomical systems in the human body	35
FAMILY	Family relation types	52
FAMILY HISTORY	Family history phrases	27
RISK FACTOR	Risk factors to health	50
NEGATION	Negative phrases	25
NO RISK FACTOR	Lack of risk factors	3
NO SYMPTOM	Lack of symptoms	2
NO COMPLAINT	Lack of complaints	1

to the corresponding entity in the knowledge base, we search for the entity with the highest cosine similarity score.

The threshold for accepting an entity is 0.8 and the entity linker uses the entity candidate with the highest score. The entity linker performs entity disambiguation. For each entity recognized by the NER component, it tries to find the associated entity in the knowledge base with the highest similarity score. As a baseline for entity linking, we perform an exact string match search in the knowledge base. We report the accuracy of the entity linker component.

3.4 Classification of the Patient Status Data

For the patient status task, the pipeline extracts several pieces of information - the basic exam results, the organ being described, and whether or not the organ is healthy/abnormal. The pipeline consists of pre-processing, extraction of exam results, organ classification, followed by abnormality classification, as shown in Fig. 2.

The pre-processing consists of sentence splitting and tokenization using NLTK library[10], abbreviation expansion using a list of predefined medical abbreviations, removing stop words using the BTB stop word list[11], transliteration of terms from Latin to Cyrillic using Python Transliterate library [12], and converting to lower case and stemming using BulStem [13].

3.4.1 Patient Status Information Extraction and Classification

The extraction of exam results is performed using regular expressions created by analyzing the Patient Status Dataset (Section 2.5).

To recognize the organ which is being described in each sentence, we use a multi-class classifier developed previously by Velichkov et al. (2020b) since it has shown almost 99% accuracy in their

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

[11] http://bultreebank.org/wp-content/uploads/2017/04/BTB-StopWordList.zip

[12] https://pypi.org/project/transliterate/

[13] https://github.com/mhardalov/bulstem-py

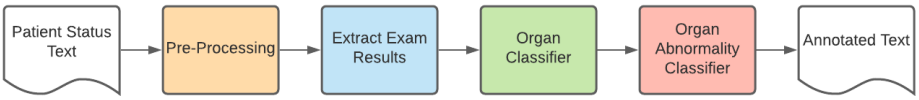

Figure 2: Architecture of the patient status pipeline.

experiments.

For the organ abnormality classification, we train a binary classifier using the Patient Status Dataset from Section 2.5. Support Vector Machine (SVM) is used to classify each record against the hypothesis "is diseased or abnormal". Based on the results, we can determine whether a particular organ is healthy or not.

The pre-processed text is transformed to feature vectors using tf-idf (TfIdfVectorizer from scikit-learn library [14]). The dataset is split into two parts - 80% for training and 20% for testing. Hyper-parameter tuning is performed using grid search. In addition, 10 fold cross-validation is used and the classification accuracy, precision, recall, and F1 scores are reported.

4 Experiments and Results

4.1 Automatic Encoding of Diagnosis into ICD-10 codes

4.1.1 Fine-Tuning MBG-ClinicalBERT Language Model

As we described above, the MBG-ClinicalBERT model is additionally trained with 10,000 articles in Bulgarian using the Bulgarian Medical Articles Dataset from Section 2.2. The dataset is split into three parts: 80% for training, 10% for hyper-parameter fine-tuning, and 10% for testing, 15% masking of the words in the text was used. In NLP the Perplexity is a measurement of how well the language model predicts a test sample. In other words, lower perplexity is better. The best result on the tuning set was achieved using the following configuration: learning rate - 4e-15, max epochs - 20. The MBG-ClinicalBERT achieved perplexity of 1.7606.

4.1.2 Fine-Tuning ICD-10 Diagnosis Classifier

For the classification task of diagnoses to ICD-10 codes, the dataset from Section 2.1 is used and was split using stratification and assigned 80% for training, 10% for hyper-parameter tuning, and 10% for testing. The best results were observed with the following configuration: learning rate - 5e-5, max epochs - 10.

The fine-tuned MBG-ClinicalBERT classifier achieved an accuracy of 92%, macro F1 of 87%, and MRR of 94%.

In order to better illustrate the task, we are solving, an example of a text and its ICD-10 classification is shown in Table 2.

The classifier is doing relatively well achieving 92% accuracy in the top 5 predictions, even though it did not predict the exact code for the sample diagnoses. The trained model can be used to suggest the top 5 codes to the doctor and have the doctor pick the most appropriate one.

For the first diagnosis, the model was able to correctly predict the 3-sign code of the disease, and the predicted 4-sign codes are quite close to the actual correct code. In the second diagnosis example, the results were slightly worse, but again all predictions are quite close to the true class. Given the presence of almost 11,000 classes, MBG-ClinicalBERT shows encouraging results for complex diagnoses.

4.2 Hybrid Method for Information Extraction from Patient History

4.2.1 Patient History Named Entity Recognition

To evaluate the NER and entity linking components, we use the Patient History Dataset from Section 2.3. A sample patient history excerpt from the test set can be seen in Table 3. A sample annotated text by the NER is displayed in Fig. 3. The symptom *"astheno-adynamia"*[15] is a complex condition of general weakness, fatigue, and helplessness that is not present in any of the dictionaries and thus cannot be recognized by the NER.

The accuracy of the NER on the test set is 87%. The majority of the errors (80%) are entities that are not recognized at all as the terms are outside of the NER dictionary. The analysis of the results of the experiments shows that the system is highly dependent on the completeness of the dictionaries it

[14] https://scikit-learn.org/stable/

[15] http://www.medik.bg/?page=test&id=367

Table 2: Example of diagnosis text and the correct and predicted ICD-10 codes.

Diagnosis Text	**Correct Code**	**Top 5 Predicted Codes**
"Неходжкинов лимфом – фоликуларен 1 ст. по СЗО, IV А кл.ст.(левкемизация, костен мозък) , FLIPI score 2 (intermediate risk), GELF score 2 (intermediate risk)." *("Non-Hodgkin's lymphoma - follicular 1 st according to WHO, IV A class (leukemia, bone marrow), FLIPI score 2 (intermediate risk), GELF score 2 (intermediate risk).")*	C82.2	C81.7 C82.1 C83 C82.9 C82
"Медикаментозно предизвикана апластична анемия." *("Drug-induced aplastic anemia.")*	D61.1	D61.9 D62 D64.0 D63 D63.8

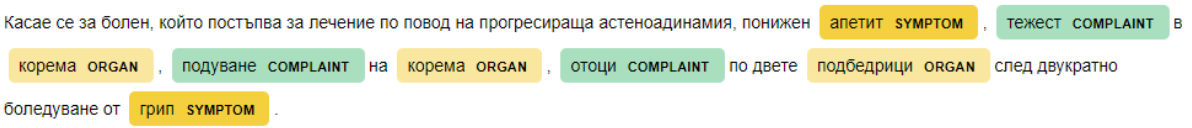

Figure 3: Sample patient history annotated by the NER.

Table 3: Sample text from the test set.

Bulgarian Text	**Translation**
Касае се за болен, който постъпва за лечение по повод на прогресираща астеноадинамия, понижен апетит, тежест в корема, подуване на корема, отоци по двете подбедрици след двукратно боледуване от грип.	*The patient was admitted for treatment with progressive astheno-adynamia, decreased appetite, heaviness in the abdomen, bloating, swelling in both lower legs after two illnesses from the flu.*

Table 4: Entities and categories predicted by the pipeline on the sample text.

Bulgarian Phrase	**Translation**	**Category**	**Correct**	**WikiData Entity ID**	**Correct**
астеноадинамия	astheno-adynamia	SYMPTOM	False	-	False
апетит	appetite	SYMPTOM	True	Q5731733	False
тежест	heaviness	COMPLAINT	True	Q104851419	False
корема	abdomen	ORGAN	True	Q9597	True
подуване	bloating	COMPLAINT	True	Q4927059	True
отоци	swelling	COMPLAINT	True	Q152234	True
подбедрици	lower leg	ORGAN	True	Q8265768	True
грип	influenza	SYMPTOM	True	Q2840	True

uses. The system can be improved by adding more entities to the dictionaries by the user. Difficulties for the system are spelling mistakes, abbreviations, and sentences, built outside the generally accepted grammatical norms.

4.2.2 Patient History Entity Linking

The entity linker produces a list of recognized entities and their respective entity IDs from WikiData as displayed in Table 4.

The overall accuracy of the Entity Linking using similarity search with MBG-ClinicalBERT embed-

Table 5: Examples of correct and incorrect organ and abnormality classification.

Sentence	Organ	Abnormal
Глава - правилна конфигурация. *(Head - normal configuration.)*	Глава *(Head)*	No
Сук реналис пол двустранно *(Sucussio renalis positive bilateral)*	Бъбрек *(Kidney)*	Yes
В добро общо състояние. *(Good overall condition.)*	None	No
Контактна, адекватна, афебрилна. *(Responsive, adequate, afebrile.)*	None	Yes

dings is 49% while the naive exact match baseline is 41%. MBG-ClinicalBERT embeddings show a statistically significant improvement over the baseline but are not high enough to be used in a real-world system. Improvements can be made using annotated data for disambiguation and fine-tuning clinical entity embeddings. Due to the huge variance in medical vocabulary as well as context-dependent abbreviations, the task of identifying the exact entity is very hard. In addition, the automatically generated translations create noise and sometimes confuse the search.

4.3 Classification of Patient Status Data

The organ abnormality classifier using an SVM model is evaluated on the Patient Status Dataset from Section 2.5. The SVM classifier shows very good results - accuracy of 93%, precision - 92%, recall - 93%, and F1 score - 92%.

Examples of the organ and abnormality classification can be viewed in Table 5. The model correctly classifies abnormality in different organs, including numerical data from basic exams, for example, sucussio renalis positive or negative. The organ abnormality classification is having issues when the sentence is describing the overall patient status using words like *"responsive"*, *"afebrile"*, which are commonly used by GPs. In these cases, the classifier incorrectly predicts the status as abnormal, even though the training dataset does not contain any abnormal samples with these words. This could be due to the short concise sentences and insufficient examples containing these words in the training dataset.

5 Conclusion and Further Work

In this paper, we presented a method for automatic extraction of data from clinical text and displaying it in a structured format for the Bulgarian language. The developed approach employs both classical methods such as those based on dictionaries and rules and modern language models based on Deep Neural Networks Transformers such as BERT. As the targeted language is a low resource one we also presented various methods for adapting the NLP tools for English and other widely spread languages. We also present results from conducted experiments to evaluate the trained ML models, which show competitive performance, making them applicable for the considered tasks.

Table 6: Results for discharge letter transformation in structure form

	Diagnosis	Patient History	Patient Status
Accuracy	92%	87%	93%
F1	87%	-	92%

Despite the presented methods are developed for a particular language they are general and can be easily adapted to other languages.

Further improvements of the presented approach can include extending the dictionaries and investigating other deep learning language models and using a semi-supervised approach to train them on the named entity recognition and text classification tasks.

Acknowledgments

This research is funded by the Bulgarian Ministry of Education and Science, grant DO1-200/2018 'Electronic health care in Bulgaria' (e-Zdrave).

Also is partially funded via GATE project by the EU Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 under GA No. 857155 and OP SE4SG under GA No. BG05M2OP001-1.003-0002-C01.

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical bert embeddings.

Svetla Boytcheva. 2011. Shallow medication extraction from hospital patient records. In *Patient Safety Informatics*, pages 119–128. IOS Press.

Svetla Boytcheva, Galia Angelova, Zhivko Angelov, and Dimitar Tcharaktchiev. 2015. Text mining and big data analytics for retrospective analysis of clinical texts from outpatient care. *Cybernetics and Information Technologies*, 15(4):58–77.

Qingxiao Chen, Xuesi Zhou, Ji Wu, and Yongsheng Zhou. 2021. Structuring electronic dental records through deep learning for a clinical decision support system. *Health Informatics Journal*, 27(1):1460458220980036.

Karl Denecke. 2008. Semantic structuring of and information extraction from medical documents using the umls. *Methods of Information in Medicine*, 47(05):425–434.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kory Kreimeyer, Matthew Foster, Abhishek Pandey, Nina Arya, Gwendolyn Halford, Sandra F Jones, Richard Forshee, Mark Walderhaug, and Taxiarchis Botsis. 2017. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *Journal of biomedical informatics*, 73:14–29.

Jianbo Lei, Buzhou Tang, Xueqin Lu, Kaihua Gao, Min Jiang, and Hua Xu. 2014. A comprehensive study of named entity recognition in chinese clinical text. *Journal of the American Medical Informatics Association*, 21(5):808–814.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics*, 9(1):1–13.

Ricky K Taira, Stephen G Soderland, and Rex M Jakobovits. 2001. Automatic structuring of radiology free-text reports. *Radiographics*, 21(1):237–245.

Dimitar Tcharaktchiev, Zhivko Angelov, Svetla Boytcheva, and Galia Angelova. 2018. Automatic generation of a national diabetes register from outpatient records. *Mathematical Modeling*, 2(4):163–166.

Nemanja Vaci, Qiang Liu, Andrey Kormilitzin, Franco De Crescenzo, Ayse Kurtulmus, Jade Harvey, Bessie O'Dell, Simeon Innocent, Anneka Tomlinson, Andrea Cipriani, et al. 2020. Natural language processing for structuring clinical text data on depression using uk-cris. *Evidence-based mental health*, 23(1):21–26.

Boris Velichkov, Simeon Gerginov, Panayot Panayotov, Sylvia Vassileva, Gerasim Velchev, Ivan Koychev, and Svetla Boytcheva. 2020a. Automatic icd-10 codes association to diagnosis: Bulgarian case. In *CSBio'20: Proceedings of the Eleventh International Conference on Computational Systems-Biology and Bioinformatics*, pages 46–53.

Boris Velichkov, Kristina Ivanova, Valeri Hristov, Ivan Borisov, Alexander Peychev, Ivan Koychev, and Svetla Boytcheva. 2020b. Ai-driven approach for automatic synthetic patient status corpus generation. In *2020 4th International Conference on Artificial Intelligence and Virtual Reality*, AIVR2020, page 29–35, New York, NY, USA. Association for Computing Machinery.

Xianghao Zhan, Marie Humbert-Droz, Pritam Mukherjee, and Olivier Gevaert. 2021. Structuring clinical text with ai: old vs. new natural language processing techniques evaluated on eight common cardiovascular diseases. *medRxiv*.

Runtong Zhang, Fuzhi Chu, Donghua Chen, and Xiaopu Shang. 2018. A text structuring method for chinese medical text based on temporal information. *International journal of environmental research and public health*, 15(3):402.