AI-driven Approach for Automatic Synthetic Patient Status Corpus Generation

Boris Velichkov
Faculty of Mathematics and
Informatics, Sofia University
"St. Kliment Ohridski", Sofia, Bulgaria
bobby.velichkov@gmail.com

Kristina Ivanova
Faculty of Mathematics and
Informatics, Sofia University
"St. Kliment Ohridski", Sofia, Bulgaria
kiivanova735@gmail.com

Valeri Hristov
Faculty of Mathematics and
Informatics, Sofia University
"St. Kliment Ohridski", Sofia, Bulgaria
valerihristov96@gmail.com

Ivan Borisov
Faculty of Mathematics and
Informatics, Sofia University
"St. Kliment Ohridski", Sofia, Bulgaria
ivbborisov@gmail.com

Alexander Peychev
Faculty of Mathematics and
Informatics, Sofia University
"St. Kliment Ohridski", Sofia, Bulgaria
a.peychev@yahoo.com

Ivan Koychev
Faculty of Mathematics and
Informatics, Sofia University
"St. Kliment Ohridski", Sofia, Bulgaria
koychev@fmi.uni-sofia.bg

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

ABSTRACT

Medical data for patients is sensitive personal information and therefore to be used in the original form is unacceptable. On the other hand, in order to be able to do various studies and analysis, we need such data. In many cases, such data even anonymized, by removing the personal identifiers, which are not suitable to be shared. Therefore we decided to create a corpus of synthetic statuses of patients that GPs place when performing a general examination. Each status consists of several sentences, each sentence describing the condition of an organ, system or part of the patient's body. We divided the status into its constituent sentences and then each sentence was classified based on the organ it refers to. We build a gold standard of manually classified sentences into list of human body organs and systems. Then we use it to train a neural network classifier of sentences that reaches almost 99% accuracy. Finally, from the all classified sentences we generate synthetic statuses, composed according to statistics in the available real statuses and medical domain constraints. The proposed approach can be easily adapted to other languages.

CCS CONCEPTS

• **Computing methodologies**; • **Natural language processing**;
• **Neural networks**; • **Machine learning**; • **Applied computing**;
• **Health informatics**;

KEYWORDS

Health Informatics, AI technologies, Big Data, Machine Learning, Text Generation, Synthetic Data, Corpus creation

ACM Reference Format:

Boris Velichkov, Kristina Ivanova, Valeri Hristov, Ivan Borisov, Alexander Peychev, Ivan Koychev, and Svetla Boytcheva. 2020. AI-driven Approach for Automatic Synthetic Patient Status Corpus Generation. In *2020 4th International Conference on Artificial Intelligence and Virtual Reality (AIVR2020)*, October 23–25, 2020, Kumamoto, Japan. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3439133.3439141>

1 MOTIVATION

In the age of digitalization and e-society, sharing data that can contribute to improving our lives is becoming increasingly valuable. Unfortunately, there is a lot of sensitive data or personal data that is not possible to be shared. The main obstacle in sharing such resources relates to ethical and legal restrictions such as GDPR legislation and privacy concerns of the users. A typical example of such data is healthcare data. In the current COVID-19 outbreak there was a huge demand on machine learning and text analytics tools that can be used to process clinical narratives, more over for non-English language. The availability of big data is crucial for training machine learning methods in healthcare.

We propose an innovative method for automatic generation of synthetic clinical notes, based on a combination of rule-based, statistical and machine learning methods. This allows generation of big open data repositories of clinical narratives, which is breakthrough

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

AIVR2020, October 23–25, 2020, Kumamoto, Japan

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-8799-6/20/10...\$15.00

<https://doi.org/10.1145/3439133.3439141>

in corpora creation in the healthcare domain. The generated data are valid and more over the synthetic corpora can contain huge variety of patient status data, which can be quite useful for modelling different phenotype and to be used in the research.

The current paper is organized as follows: Section 2 overviews approaches for synthetic data generation in Healthcare; Section 3 describes the real data and their preprocessing, on the top of which are generated the synthetic data; Section 4 presents in detail the proposed AI-based approach for automatic generation of synthetic patient status data; and Section 5 concludes and sketches some directions for further work.

2 RELATED WORK

Various methods for synthetic data generation have been developed. The most important and desired feature of such repositories is the validity of data and minimization of the information loss.

The problem with the availability and access of the research data in the healthcare sector is addressed by Shamsuddin, Rittika, et al [10]. They also offer a Virtual Patient Model (VPM), a process that combines the use of an optimization algorithm, statistical analysis, and machine learning techniques to generate synthetic time series data and report their effectiveness in predictive models. The proposed model is validated by implementing a genetic algorithm that captures important characteristics of real-time patient data (original). As a result of applying constraints in the generative process, the algorithm outputs a best fit synthetic candidate solution of the original time series data. Experimental results using statistical tests to verify both the synthetic and original datasets show that the synthetic dataset retains features from the original dataset. The authors of the paper claim that by using machine learning prediction models in combination with synthetic data as a training dataset, they manage to achieve promising results in the class recognition compared to the training with the original data. The first simulation framework which allows the generation of realistic 3D synthetic cardiac US and MR (both cine and tagging) image sequences from the same virtual patient is proposed by Zhou, Yitian, et al [14]. The approach for VPM repositories generation is explored in [6] as well.

There are several attempts for generation of realistic synthetic Electronic Health Records (EHRs) [7], [5], [3]. Walonoski, Jason, et al [12] developed Synthea, an open-source software package that simulates the lifespans of synthetic patients, modeling the 10 most frequent reasons for primary care encounters and the 10 chronic conditions with the highest morbidity in the United States. It is expected that by engaging a growing community of users, the synthetic data generated will become increasingly comprehensive, detailed, and realistic over time. Also the authors mention that synthetic patients can be simulated with models of disease progression and corresponding standards of care to produce risk-free realistic synthetic health care records at scale. Zand, Ramin, et al [13] describe the importance of building synthetic patient populations and customized, predictive models using heterogeneous datasets, such as electronic health records (EHRs). This is a pioneering work on the use of in silico clinical trials to accelerate the development of new drugs.

Another important aspect that was addressed by Chen et al [4] is the validity of synthetic clinical data and formal methods for assessment.

Baytas, Inci M., et al [1] propose a novel LSTM unit called Time-Aware LSTM (T-LSTM) to handle irregular time intervals in longitudinal patient records. In their paper they claim that experiments on synthetic and real world datasets show that the proposed T-LSTM architecture captures the underlying structures in the sequences with time irregularities. Begoli et al [2] propose a method for synthetic generation of mental health characteristics reports, based on deep learning generative methods applied over literature and public statistical models. Several applications have already been developed using synthetic data corpora, for example development of family history annotation guidelines presented by Rama et al [9].

We focus on only a small part of the outpatient record generation, namely the patient status. The proposed approach is based on a natural language generation method using statistical and rule-based templates.

3 DATA

Clinical patient data cannot be used for research, due to ethical and legal restrictions, because they contain sensitive personal information. Even anonymizing patient data does not guarantee their security. There are cases in which anonymized data is de-identified.

The data preparation task consists of three subtasks (see Figure 1): (1) repository creation from real data; (2) patient status template modelling and (3) preparation of small training corpus for classification of sentences according to anatomical organs and systems they describe.

3.1 Repository Creation

For this research we used a repository of real anonymized outpatient records (ORs) described by General Practitioners (GPs), from the Bulgarian National Diabetes Register [11]. In order to prevent any data breach the ORs were splitted on sentences and was generated repository of individual sentences used in real ORs. In this study we focused on sentences that are used to describe patient status only.

Each patient status description consists of several sentences, each sentence describing the condition of an anatomical organ, system or part of the patient's body. A patient status description can range from 4-5 to 12-13 sentences, in which some anatomical organs almost always present in ORs, whereas other organs are less frequently described.

The first subtask is the repository generation from the real anonymized ORs. One of the most complex and challenging tasks is to process the real ORs - namely the sentence splitting of the patient status data. The difficulty in this case comes from the fact that in the sentences themselves there are used many abbreviations. The patient status presentation is more in telegraphic style, rather than using complete sentences, which implies a lack of punctuation. This makes almost impossible to determine where the description of the condition of one anatomical organ ends and where the next one begins.

Figure 1: Data preparation: (1) repository creation from real data; (2) patient status template modelling and (3) preparation of small training corpus for classification of sentences according to anatomical organs and systems they describe.

Figure 2: Synthetic patient status data template.

Some additional specific sentence splitting rules were introduced. For instance, in the sentence "Копем- б.о. Черендробислезка- б.о. Succ. ren: -, Периференилимфнивъзлинисепалпиратувеличени" (*"Abdomen-w.c. Liver and spleen-w.c. succ.ren: -, Peripheral lymph nodes-Not palpable"*), the phrase "б.о." (*"without complications"*) is a signaling for the end of an anatomical organ description and can be used for sentence splitting rule.

After splitting the patient status by sentences from 10,000 patients ORs we generated a repository of about 1 million individual sentences. From the generated repository were removed duplicated sentences, by preserving information about the frequency. The result repository contain 100,000 unique sentences.

3.2 Synthetic Patient Status Data Templates Modelling

The next subtask is templates modelling for the synthetic patient status data. The real anonymized ORs were used for desk research only in modelling the templates. Based on the frequency analysis were modelled different patient status description patterns (see Figure 2).

3.3 Training Corpus with Categorized Patient Status Descriptions

One of the main subtasks in synthetic data generation is to train text-based classification method for individual sentences classification to categories representing different anatomical organ, systems and parts of human body.

Since there is lack of language resources for medical domain in Bulgarian language, the initial step was to create training corpus. Some representative subsets of different sentences were selected from the generated repository. The data cleaning was applied and then the result set of clear sentences was labeled manually. Web based data curation tool was developed to assist this activity.

For this initial subtask were used the word frequency in the repository. The most commonly described anatomical organs are such as heart, lungs, liver, limbs, skin, eyes, nose, throat, nervous system, etc. Another important feature is the use of many Latin terms in the Bulgarian clinical narratives. The Bulgarian medical terminology is a mixture between medical terms in Bulgarian, medical terms in Latin and Latin medical terms transliteration in Cyrillic. Adding to this the abbreviations for all three cases, categorizing an anatomical organ becomes a challenging task.

The real data contain a lot of noise - typos, concatenated words, homographs, etc. Some preliminary cleaning of the repository was performed. A balanced training set for all anatomical organs and systems was selected, where only sentences representing information for single anatomical organ/system were selected in the training corpus in order to reduce the ambiguity.

The created training corpus contains 17,000 manually classified sentences to 25 different organs and systems. Some examples of sentences that refer to only one organ can be seen in Table 1. In this case the examples are for heart and lung.

Table 1: An example of sentences that refer to only one organ (in this case the organs are heart and lung)

Anatomical Organ/ System	Sentence
Lung	БЯЛДРОБ–чистовезикуларнодишане, безхрипованаходка(<i>LUNGS - pure vesicular respiration, no wheezing</i>)
Lung	Пулмо: отслабеновезикуларнодишане(<i>Pulmo: weakened vesicular respiration</i>)
Lung	Пулмо–вез.дишане, судълженоиздишване, ед.сухисв.хрипове(<i>Pulmo - weighted breathing, prolonged exhalation, single dry wheezing</i>)
Heart	СЪРЦЕ–ритмичнасърдечнадейност, глухитонове(<i>HEART - rhythmic heart activity, deaf tones</i>)
Heart	СЪРЦЕ–Б.О(<i>HEART- without complaints</i>)
Heart	КОР–РСД(<i>COR- rhythmic heart activity</i>)

Figure 3: Synthetic patient status data generation.

4 METHODS

The synthetic corpus generation is based on statistical models, that use the templates created over real patient status data, and text-based classification methods applied on the preprocessed repository of sentences.

4.1 Sentence Preprocessing

The division of the statuses into separate sentences appears to be a quite complex task. The difficulty in this case comes from the fact that in the sentences themselves there are many abbreviations of words, different abbreviations of organs and systems are used, sometimes no punctuation marks and it becomes almost impossible to determine where the description of the condition of an organ ends and where the next one begins. This subtask contains the classical preprocessing pipeline steps like tokenization, text normalization, stop words removal, stemming, etc.

Tokenization: As a token we consider words only. Thus all numerical values are skipped, as well as white spaces, and punctuation. The omitted data do not bring any additional value to the task of anatomic organ/system classification.

Text normalization: Since the outpatient record is created during the patient's examination, general practitioners usually use plenty of abbreviations and acronyms due to the time shortage. We have compiled a list of abbreviation and their corresponding

extended meaning. This is an essential step of the sentence categorization because very often the abbreviation of a system of organs consists of the organ to which we want to categorize the sentence.

For example: "CCC" (cardio vascular system)

- "Сърдечносъдовасистема" (cardio vascular system)
- "Сърдечно" (cardiac) has the same root as the word (heart) "сърце" which can help in the comparison step.

Stop words removal: This step brought some benefits on the performance. As a resource was used available list of stop word for Bulgarian language provided by Bultreebank¹. As it left fewer but only meaningful tokens, it decreased the size of the dataset which resulted in speeding up the classification process and increasing the accuracy.

Examples:

- (abdomen without features) Копемб.о. → Копембо → Копем
- Sentences describing patient's general and mental condition: These sentences are ignored for categorization, later on, on synthetic status generation they are added with the respective probability of encounters.

Stemmig: Each word of the sentence is stemmed by inflectional stemmer for Bulgarian BulStem [8]. All the available stemming rules are imported but due to the specificity and terminology of

¹<http://bultreebank.org/bg/resources/>

Table 2: An example of sentences from some of the clusters

Cluster	Sentences
General Status	Общосъстояние- Доброобщосъстояние, Безпромянаобщотосъстояние(General condition - Good general condition, No change in general condition)
Body Regions	Регионинатялото- ГЛАВАиШИЯ-Безособености, КОРЕМ-Мек, палпаторнонеболезнен, Крайници-безототи(Body Regions - HEAD and NECK- without complications, Abdomen - Soft, palpably painless, Limbs - no swelling)
Respiratory system	Дихателнасистема- БЯЛДРОБ-чистовезикуларнодишане, Пулмо-везикуларнодишанебезхрипове, Гърло: Нормално, сливици-безаномалии(Respiratory system - LUNGS-pure vesicular respiration, Pulmonary-vesicular respiration without wheezing, Throat: Normal, tonsils-without abnormalities)
Cardio Vascular system	Сърдечносъдовасистема- СЪРЦЕ-ритмичнасърдечнадейност, Ритмичнасърдечнадейност- яснитонове, безприбавенишумове, Cor - р.с.д., fr 72 умин. , RR 130 80 (Cardiovascular system - HEART-rhythmic heartbeat, Rhythmic heartbeat - clear tones, without added noise, Cor - rhythmic heartbeat, fr 72 HR , RR 130 80)
Digestive system	Храносмилателнасистема- ЧЕР.ДРОБИСЛЕЗКА-Неувеличени, ЕЗИК-Влаженнеобложен, Черендробислезканесепалпират(Digestive system - LIVER AND SPLEEN-Not enlarged, TONGUE-Moist uncoated, Liver and spleen are not palpable)
Haemic and Immune Systems	ХемичнаиИмуннасистема- ПЕРИФЕРНИЛИМФНИВЪЗЛИ-Неувеличени, Далак: Несепалпира, Слезка-неувеличена(Haemic and Immune System - PERIPHERAL LYMPH -Not enlarged, Spleen: Not palpable, Spleen – not enlarged)

the medical terminology and the prevalence of Latin terms, the stemmer's rules are not enough, and a small amount of the words is actually stemmed. Moreover, in some cases the word's length goes nearly two or three characters which results in one more step in the preprocessing, because words of this type are not eligible for comparison.

Examples: (afebrile) афебрилен → афебрилен

4.2 Classification of Sentences into Human-body Organs and Systems

4.2.1 Lexicon Based Classification. The available data about organs and systems is held in two main resources:

- **GPs lexicon** with an anatomical organs/system, which are most common in general practitioner's examinations in our resource bank.
- **Human body structure dictionary** with most of the organs in the human body, according to the medical terminology standard classifications.

This allows searching also for categories descriptions that are out of the vocabulary of our ORs repository.

Searching in GPs lexicon: The searching is happening using TF, (term frequency) as this data is manually extracted, the results that TF is giving are sufficient. Very often the general practitioners are describing organs which are semantically connected as liver and spleen. Because of that we have a list with most of the described pairs of anatomical organs in the ORs. After choosing the most appropriate category candidate then in a list of anatomical organ pairs is searched. If a result is found, then the sentence is annotated with the pair match else the initial organ becomes the sentence's category.

4.2.2 Searching in The Human Body Structure Dictionary. If the searching in GPs lexicon does not return a result, then in the list of all organs is searched.

Figure 4: LSTM training over 17 thousand sentences divided into 25 classes(organs), in 5 epochs. The achieved accuracy is about 98.9%.

In order to find an organ there a slight customization of cosine metric is done - if a potential category is an n-gram and it's contained in the sentence then a maximal score is given, else the standard implementation is applied. The category with the highest score is checked in the pairs list and then used for final category.

The ignored sentences in the preprocessing step about the mental and general state of the patient are added with the same frequency of encounter and annotated with default category "General Status", so later on they can be added in the status generation.

4.2.3 LSTM Multi-Class Classification. From the gold standard of manually classified sentences, we decided to train a neural network to be used in the further classification of sentences and to compare the results obtained. For this purpose, we used the LSTM model of a recurrent neural network for multiple classification. For the realization of the recurrent neural network we used Python and Keras, using the following parameters - activation function="softmax" and optimizer="Adam". Because it is a multi-class classification task,

Table 3: Table with the specific number of sentences, unsuitable for classification by organ category and description of the reason. Below they are taken out of the gold standard and the categorization algorithm is rated on this set as well

Attribute	Count	Consideration
blood pressure	1722	Organ description/status
weight	163	Organ description/status
height	104	Organ description/status
temperature	172	Organ description/status
lymph nodes	193	Organ pair
pulse	355	Organ description/status
head	271	Organ pair

Table 4: Table with classification results of the manually annotated list of sentences to only comparable ones obtained by taking in account the described above considerations

#	All sentences	Comparable Sentences
Number of sentences	12404	9015
Correct classified (count)	6045	6045
Incorrect classified (count)	6350	2970
Correct classified (%)	48.73%	67.05%
Incorrect classified (%)	51.26%	32.95%

categorical crossentropy is used as the loss function. By submitting the gold standard of about 17 thousand sentences divided into 25 classes (organs), in 5 epochs of training we achieved about 98.9% accuracy (see Figure 4).

4.2.4 Algorithm Evaluation. The categorization algorithm is evaluated against 12 404 manually annotated sentences.

4.2.5 Considerations. In the synthetic status generation and data preparation steps, some categories are divided into smaller pieces and follow the exemplary hierarchy: Cardiovascular system → heart → blood pressure, pulse

As pulse and blood pressure are not organs but a description of such, this leads to false negative results in the comparison. It's the same for sentences which have the description of two organs at the same time like lymph nodes and thyroid gland. Also, the organ dataset is referring to head and neck as one organ, but the manual annotation is referring to them as two - separate ones.

4.2.6 Evaluation. Differences between the manually annotated and the algorithm result are present in a total of 6359 sentences, which are exactly 51.27%.

If we have the considerations in account then the incorrect classified sentences are 2970 in total, for more details see Table 3 and Table 4. As the analysis of the false negative categorization is done manually it is possible for some cases to be missing.

4.3 Synthetic Data Generation

After successful classification of sentences to anatomical organs, pair of organs and systems, comes the generation of the synthetic patient status corpora (see Figure 3). Since sentences are categorized to organs, pair of organs or systems we iterate and split them in clusters of sentences each cluster representing higher level human body systems which are the constituents of an ambulatory status. As mentioned in the data processing section we identified a not so large number of organs and systems that present in the description of the statuses from the whole corpus, created a list of them and identified the human body systems that each organ belongs to. The reorganization of sentences into different clusters was realized using already categorized sentences to organs and the map of organ to system. Even though we have relatively big list of human body systems, only 11 of them will take part in the templates for generation as the other have zero sentences categorized to them. On the one hand the reorganization of sentences into clusters will enable us to execute frequency analysis of the sentences in each cluster, on the other we follow Stratified Random Sampling approach which requires to group sentences and skip overlapping. Some examples of sentences from some of the clusters are shown in Table 2.

Since there are guidelines that general practitioners follow, but they are not strict giving some freedom to the doctor, we decided to build a standard for our synthetic statuses. A status will consist of no more than 11 sentences, each referring to unique human body system, ordered in a predefined way generally followed by doctors. The ordered categories are: "General status", "Body regions", "Respiratory System", "Cardiovascular System", "Digestive System", "Urogenital System", "Endocrine System", "Nervous System", "Sense Organs", "Musculoskeletal System", "Haemic and Immune Systems". Some statuses can consist of the maximum number of sentences which in our case is 11, but most of the statuses will consist of 7, 8 sentences or even less, for this purpose systems are split into obligatory and optional sections and each optional section has assigned probability, based on how often it is included by the doctors in the original statuses. For example, the first section of a status is "General status" which is obligatory and always present, as opposed to other optional categories like "Sense Organs" which can be found only in 10percent of the generated synthetic statuses. The number of sentences, respectively the number of human body systems in a status is dynamic, it always includes the obligatory systems: "General status", "Respiratory System", "Cardiovascular System", "Digestive System", the rest of them are optional sections and the number of optional sections included in a status is generated at random. In Table 5 are shown some examples of synthetic statuses, one that includes all sections and one that includes only a few.

5 CONCLUSION AND FURTHER WORK

In this paper we proposed a method for generating a corpus of synthetic statuses of patients that GPs place when performing a general examination. The approach uses a corpus of sentences taken from real statuses, where each sentence describing the condition of an organ, system or part of the patient's body. We divided the status into its constituent sentences and then each sentence was classified based on the organ it refers to. We develop an approach for division of the statuses into separate sentences, which appears to be to be a

Table 5: Examples of synthetically generated patient statuses, one that includes all sections and some that includes optional sections

#	Status with obligatory with 2 optional sections and with all optional sections
1	Вдоброобщо състояние. ГЪРЛО-Безособености. СЪРЦЕ-ритмична сърд. дейност. ЕЗИК-Влаженно обложен. (In good overall condition. THROAT-No features. HEART-rhythmic heart activity. LANGUAGE-Wet uncoated.)
2	Доброобщо състояние. БЯЛДРОБ-Леко отслабено везикуларно дишане. Сър - р.с.д., Черендробислезка не са палпиратувеличени. ОДА- б.о. ПЕРИФЕРНИ ЛИМФНИ ВЪЗЛИ-Неувеличени. (In good overall condition. LUNGS-Slightly weakened vesicular respiration. Cor - rhythmic heartbeat, Liver and spleen are not palpated enlarged. Musculoskeletal System - without complaints PERIPHERAL LYMPH -Not enlarged.)
3	Тегло 72 кг. КОРЕМ-Мек палпаторно не болен. БЯЛДРОБ-чисто везикуларно дишане. ПУЛС-68. Черендробислезка неувеличени. Сук. реналис отр. Щитовидна жлезка неувеличена. Нервна система: Без невролог. Зрение- запазено. ОПОРНО-ДВИГ. АПАРАТ-болка в двете коленни стави.. Лигавици бледорозови. ПЛВ не са палпиратувеличени. (Weight 72kg. ABDOMEN - Soft palpably painless. LUNGS - pure vesicular respiration. Heart rate - 68. Liver and spleen - not enlarged. Succusio renalis negative. Thyroid gland not enlarged Nervous system: Without a neurologist. Vision-preserved. MUSCULOSKELETAL SYSTEM-pain in both knee joints. Mucous membranes pale pink PERIPHERAL LYMPH are not palpated enlarged.)

quite challenging task. The approach proposes techniques to deal with: abbreviations of words; abbreviations of organs and systems; missing punctuation marks, which makes almost impossible to determine where the description of the condition of an organ ends and where the next one begins. We build a gold standard of manually classified sentences into list of human body organs and systems. Then we use it to train a neural network classifier of sentences that reaches almost 99% accuracy. Finally, from all classified sentences we generate synthetic statuses, composed according to statistics in the available real statuses taking in to account medical domain constraints. The proposed approach is as possible language independent by design and can be easily adapted to other languages. As further works we are planning to extend our corpus and to explore the applicability of recent language model such as Bert for generation of synthetic patient statuses.

ACKNOWLEDGMENTS

This research is partially funded by the Bulgarian National Science Fund, grant DN-02/4-2016 'Specialized Data Mining Methods Based on Semantic Attributes' (IZIDA). We also acknowledge the provided access to the infrastructure of the Centre for Advanced Computing and Data Processing, with the financial support by the Grant No BG05M2OP001-1.001-0003, financed by the Science and Education for Smart Growth Operational Program (2014–2020) and co-financed by the European Union through the European structural and investment funds.

REFERENCES

- [1] Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. 2017. Patient subtyping via time-aware LSTM networks. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 65–74.
- [2] Edmon Begoli, Kris Brown, Sudarshan Srinivas, and Suzanne Tamang. 2018. SynthNotes: A Generator Framework for High-volume, High-fidelity Synthetic

Mental Health Notes. In 2018 IEEE International Conference on Big Data (Big Data). IEEE, 951–958.

- [3] Anna L Buczak, Steven Babin, and Linda Moniz. 2010. Data-driven approach for creating synthetic electronic medical records. BMC medical informatics and decision making 10, 1 (2010), 59.
- [4] Junqiao Chen, David Chun, Mitesh Patel, Epton Chiang, and Jesse James. 2019. The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. BMC medical informatics and decision making 19, 1 (2019), 44.
- [5] Kudakwashe Dube and Thomas Gallagher. 2013. Approach and method for generating realistic synthetic electronic healthcare records for secondary use. In International Symposium on Foundations of Health Informatics Engineering and Systems. Springer, 69–86.
- [6] Uri Kartoun. 2016. A methodology to generate virtual patient repositories. arXiv preprint arXiv:1608.00570 (2016).
- [7] Scott H Lee. 2018. Natural language generation for electronic health records. NPJ digital medicine 1, 1 (2018), 1–7.
- [8] Preslav Nakov. 2003. BulStem: Design and evaluation of inflectional stemmer for Bulgarian. In Workshop on Balkan Language Resources and Tools (Balkan Conference in Informatics).
- [9] Taraka Rama, Pål Brekke, Øystein Nytrø, and Lilja Øvreliid. 2018. Iterative development of family history annotation guidelines using a synthetic corpus of clinical text. In Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis. 111–121.
- [10] Rittika Shamsuddin, Barbara M Maweu, Ming Li, and Balakrishnan Prabhakaran. 2018. Virtual patient model: an approach for generating synthetic healthcare time series data. In 2018 IEEE International Conference on Healthcare Informatics (ICHI). IEEE, 208–218.
- [11] Dimitar Tcharaktchiev, Zhivko Angelov, Svetla Boytcheva, and Galia Angelova. 2018. Automatic generation of a national diabetes register from outpatient records. Mathematical Modeling 2, 4 (2018), 163–166.
- [12] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. 2018. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. Journal of the American Medical Informatics Association 25, 3 (2018), 230–238.
- [13] Ramin Zand, Vida Abedi, Raquel Hontecillas, Pinyi Lu, Nariman Noorbakhsh-Sabet, Meghna Verma, Andrew Leber, Nuria Tubau-Juni, and Josep Bassaganya-Riera. 2018. Development of synthetic patient populations and in silico clinical trials. In Accelerated Path to Cures. Springer, 57–77.
- [14] Yitian Zhou, Sophie Giffard-Roisin, Mathieu De Craene, Sorina Camarasu-Pop, Jan D'Hooge, Martino Alessandrini, Denis Friboulet, Maxime Sermesant, and Olivier Bernard. 2017. A framework for the generation of realistic synthetic cardiac ultrasound and magnetic resonance imaging sequences from the same virtual patients. IEEE transactions on medical imaging 37, 3 (2017), 741–754.