Clinical Text Mining for Context Sequences Identification

Svetla Boytcheva

Clinical Text Mining for Context Sequences Identification

Svetla Boytcheva[0000−0002−5542−9168]

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

Abstract. This paper presents an approach based on sequence mining for identification of context models of diseases described by different medical specialists in clinical text. Clinical narratives contain rich medical terminology, specific abbreviations, and various numerical values. Usually raw clinical texts contain too many typos. Due to the telegraphic style of the text and incomplete sentences, the general part of speech taggers and syntax parsers are not efficient in text processing of non-English clinical text. The proposed approach is language independent. Thus, the method is suitable for processing clinical texts in low resource languages. The experiments are done on pseudonimized outpatient records in Bulgarian language produced by four different specialists for the same cohort of patients suffering from similar disorders. The results show that from the clinical documents can be identified the specialty of the physician. Even the close vocabulary is used in the patient status description there are slight differences in the language used by different physicians. The depth and the details of the description allow to determine different aspects and to identify the focus in the text. The proposed data driven approach will help for automatic clinical text classification depending on the specialty of the physician who wrote the document. The experimental results show high precision and recall in classification task for all classes of specialist represented in the dataset. The comparison of the proposed method with bag of words method show some improvement of the results in document classification task.

Keywords: Data Mining · Text Mining · Health Informatics.

1 Motivation

Healthcare is data intensive domain. Large amount of patient data are generated on daily base. However, more than 80% of this information is stored in non structured format - as clinical texts. Usually clinical narratives contain description with sentences in telegraphic style, non-unified abbreviation, many typos, lack of punctuation, concatenated words, etc. It is not straightforward how patient data can be extracted in structured format from such messy data. Natural language processing (NLP) of non-English clinical text is quite challenging task due to lack of resources and NLP tools [11]. There are still non existing translations of

SNOMED[1], Medical Subject Headings (MeSH)[2] and Unified Medical Language System (UMLS)[3] for the majority of languages.

Clinical texts contain complex descriptions of events. Investigating the cumulative result of all events over the patient status require more detailed study of different ways of their description. All physicians use common vocabulary and terminology to describe organs and systems during the human body observation but tend to use different description depending on their specialty. Analyzing complex relations between clinical events will help to prove different hypothesis in healthcare and automatically to generate context models for patient status associated to diagnoses. This is very important in epidemiology and will help monitoring some chronic diseases' complications on different stages of their development. The chronic disease with highest prevalence are cardiovascular diseases, cancer, chronic respiratory diseases and diabetes[4]. The complications of these chronic diseases develop over time and they are with high socioeconomic impact and the main reason for over than 70% of mortality cases. In this paper are presented some results for processing data of patients with Diabetes Mellitus type 2 (T2DM), Schizophrenia (SCH) and Chronic Obstructive Pulmonary Disease (COPD).

We show that data mining and text mining are efficient techniques for identification of complex relations in clinical text.

The main goal of this research is to examine differences and specificity in patient status description produced by different medical specialists. The proposed data-driven approach is used for automatic generation of context models for patient status associated with some chronic diseases. The approach is language independent. An application of the context sequences in used for clinical text classification depending on the specialty of the physician who wrote the document.

The paper is structured as follows: Section 2 briefly overviews the research in the area; Section 3 describes the data collections of clinical text used in the experiments; Section 4 presents the theoretical background and formal presentation of the problem; Section 5 describes in details the proposed data mining method for context models generation from clinical text; Section 6 shows experimental results and discusses the method application in clinical texts classification; Section 7 contains the conclusion and sketches some plans for future work.

2 Related Work

Data mining methods are widely used in clinical data analyses both for structured data and free text [17]. There are two types of frequent patterns mining frequent itemsets patterns mining (FPM) and frequent sequence mining (FSM).

[1] SNOMED, https://www.snomed.org/
[2] Medical Subject Headings MESH, https://www.nlm.nih.gov/mesh/
[3] UMLS, https://www.nlm.nih.gov/research/umls/
[4] World Health Organization (WHO) fact sheets, http://www.who.int/mediacentre/factsheets/fs355/en/

In the first approach the order of items does not matter, and in the second one the order does matter.

In context modeling task there is some research for other domains. Ziembiski [19] proposes a method that initially generates context models from small collections of data and later summarizes them in more general models. Rabatel et al [14] describes a method for mining sequential patterns in marketing domain taking into account not only the transactions that have been made but also various attributes associated with customers, like age, gender and etc. They initially uses classical data mining method for structured data and later is added context information exploring the attributes with hierarchical organization.

Context models in FPM are usually based on some ontologies. Huang et al [7] present two semantics-driven FPM algorithms for adverse drug effects prevention and prediction by processing Electronic Health Records (EHR) The first algorithm is based on EHR domain ontologies and semantic data annotation with metadata. The second algorithm uses semantic hypergraph-based k-itemset generation. Jensen et al [8] describe a method for free text in Electronic Health Records (EHR) processing in Norwegian language. They are using NOR-MeSH for estimation of disease trajectories of the cancer patients.

One of the major problems with clinical data repositories is that they contain in-complete data about the patient history. Another problem is that the raw data are too noisy and needs significant efforts for preprocessing and cleaning. The timestamps of the events are uncertain, because the physicians dont know the exact occurrence time of some events. There can be a significant gap between the onset of some dis-eases and the first record for diagnosis in EHR made by the physician. Thus a FPM method for dealing with temporal uncertainty was proposed by Ge et al [4]. It is hard to select representative small collections of clinical narratives, because there is a huge diversity of patient status descriptions. Some approaches use frequent patterns mining (FPM) considering the text as bag-of-words and losing all grammatical information.

The majority of FSM and FPM applications in Health informatics are for patterns identification in structured data. Wright et al [16] present a method for prediction of the next prescribed drug in patient treatment. They use CSPADE algorithm for FSM of diabetes medication prescriptions. Patniak et al [12] present mining system called EMRView for identifying and visualizing partial order information from EHR, more particularly ICD-10 codes.

But there are also applications of FSM for textual data. Plantevit et al [13] present a method for FSM for Biomedical named entity recognition task.

There are developed a variety of techniques for FPM and FSM task solution. Some of them are temporal abstraction approach for medical temporal patterns discovery, one-sided constitutional nonnegative matrix factorization, and symbolic aggregate approximation [15].

Healthcare is considered as data-intensive domain and as such faces the challenges of big data processing problems. Krumholz [10] discusses the potential and importance of harnessing big data in healthcare for prediction, prevention and improvement of healthcare decision making.

In the classification task there are used successfully many artificial intelligence (AI) approaches [9] with high accuracy: neural networks, naive Baise classifiers, support vector machines, etc. The main reason for choosing FSM method is than in healthcare data processing the most important feature of the used method is the result to be explainable, e.i. so called "Explainable AI" [5]. This will make the decision making process more transparent.

3 Materials

For experiments is used a data collections of outpatient records (ORs) from Bulgarian National Diabetes Register [2].

They are generated from a data repository of about 262 million pseudonimized outpatient records (ORs) submitted to the Bulgarian National Health Insurance Fund (NHIF) in period 2010–2016 for more than 5 million citizens yearly. The NHIF collects for reimbursement purpose all ORs produced by General Practitioners and the Specialists from Ambulatory Care for every patient clinical visit. The NHIF collects for reimbursement purpose all ORs produced by General Practitioners and the Specialists from Ambulatory Care for every patient clinical visit. The collections used for experiments contain ORs produced by the following specialists: Otolaryngology (S14), Pulmology (S19), Endocrinology (S05), and General Practitioners (S00).

ORs are stored in the repository as semi-structured files with predefined XML-format. Structured information describe the necessary data for health management like visit date and time; pseudonimized personal data and visit-related information, demographic data (age, gender, and demographic region), etc. All diagnoses are presented by ICD–10[5] codes and the name according to the standard nomenclature. The most important information concerning patient status and case history is provided like free text.

For all experiments are used raw ORs, without any preprocessing due to the lack of resources and annotated corpora. The text style for unstructured information is telegraphic. Usually with no punctuation and a lot of noise (some words are concatenated; there are many typos, syntax errors, etc.). The Bulgarian ORs contain medical terminology both in Latin and Bulgarian. Some of the Latin terminology is also used with Cyrillic transcription.

The most important information concerning patient status and case history is provided like free text. ORs contain paragraphs of unstructured text provided as separate XML tags (see Table 1): Anamnesis, Status, Clinical tests, and Prescribed treatment.

4 Theoretical Background

Lets consider each patient clinic visits (i.e. OR) as a single event. For the collection S we extract the set of all different events. Let $E = \{e_1, e_2, \ldots, e_k\}$ be

[5] http://apps.who.int/classifications/icd10/browse/2016/en#/

XML field	Content
Anamnesis	Disease history, previous treatments, family history, risk factors
Status	Patient state, height, weight, BMI, blood pressure etc.
Clinical tests	Values of clinical examinations and lab data listed in arbitrary order
Prescribed treatment	Codes of drugs reimbursed by NHIF, free text descriptions of other drugs and dietary recommendations

Table 1. Fields with free text in ORs that supply data for data mining components

the set of all possible patient events. The vocabulary $W = \{w_1, w_2, \ldots, w_n\}$, used in all events E in S will be called *items*, where $e_i \subseteq W,\ 1 \leq i \leq N$. Lets $P = \{p_1, p_2, \ldots, p_N\}$ be the set of all different patient identifiers in S. The associated unique transaction identifiers (tids) shall be called *pids* (*patient identifiers*).

Let each sentence in a clinical text e_1 is splitted on a sequence of tokens $X \subseteq W$. X is called an *itemset*. For each itemset X is generated a vector (sequence) $v = \langle v_1, v_2, \ldots, v_m \rangle$, where $v_i \in W,\ 1 \leq i \leq N$. The length of a sequence v is m (the number of tokens), denoted $len(v) = m$. We denote $\emptyset$ the empty sequence (with length zero, i. e. $len(\emptyset) = m$).

Let $D \subseteq P \times E$ be the set of all sequences in collection in the format $\langle pid, sequence \rangle$. We will call D *database*.

Let $p = \langle p_1, p_2, \ldots, p_m \rangle$ and $q = \langle q_1, q_2, \ldots, q_t \rangle$ be two sequences over W. We say that q is *subsequence of* p denoted by $q \subseteq p$, if there exists one-to-one mapping: $\theta : [1, t] \rightarrow [1, m]$, such that $q_i = p_{\theta(i)}$ and for any two positions i, and j in q, $i < j \Rightarrow \theta(i) < \theta(j)$.

Each sequential pattern is a sequence. A sequence $A = X_1, X_2, \ldots, X_m$, where $X_1, X_2, \ldots, X_m$ are itemsets is said to *occur in another sequence* $B = Y_1, Y_2, \ldots, Y_t$, where $Y_1, Y_2, \ldots, Y_t$ are itemsets, if and only if there exist integers $1 \leq i1 < i2 \ldots < ik \leq t$ such that $X_1 \subseteq Y_{i1}, X_2 \subseteq Y_{i2}, \ldots, X_m \subseteq Y_{im}$.

Let D is a database and $Q \subseteq E$ is a sequential pattern. The *support* of a sequential pattern Q, denoted $support(Q)$ is the number of sequences where the pattern occurs divided by the total number of sequences in the database.

We define minimal support threshold *minsup* - a real number in the range [0,1]. A frequent sequential pattern is a sequential pattern having a support no less than *minsup*.

In our task we are looking only for frequent sequential pattern for given *minsup*.

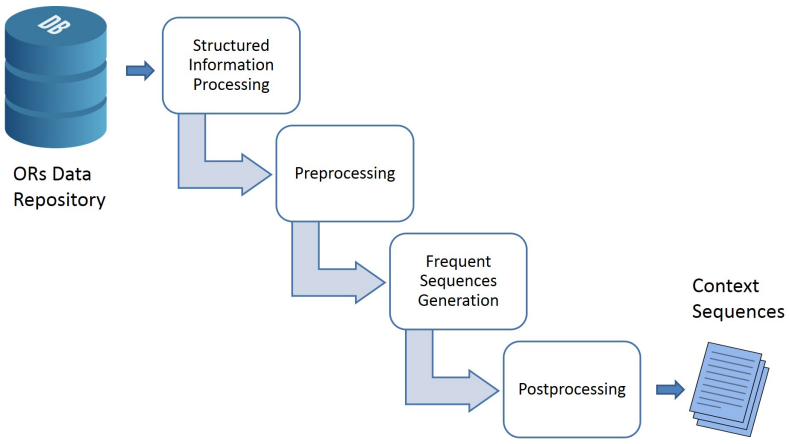

Fig. 1. Pipeline for automatic context Sequences generation

5 Method

Initially we generate collections $S_1, S_2, \ldots, S_r$ of ORs from the repository, using the structured information data for specialists who wrote them. We define vocabularies $W_1, W_2, \ldots, W_r$.

The collections processing is organized as pipeline (see Fig.1). The first step is to split each collection on two subsets - one that contain only Anamnesis for patients (SA_i) and the other SH_i - for their Status. Each of these subsets will be processed independently. We define for all collections vocabularies $WA_1, WA_2, \ldots, WA_r$ and $WH_1, WH_2, \ldots, WH_r$ for each or these subsets correspondingly.

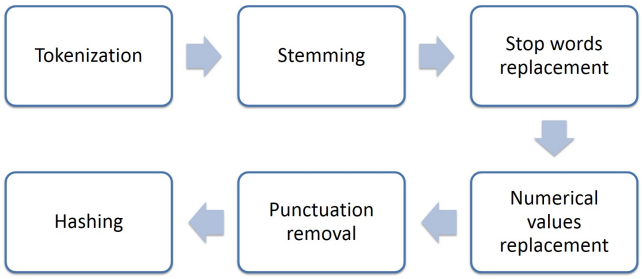

Fig. 2. Pipeline for preprocessing of free-text in outpatient record

The next step converts free text from ORs into database (see Fig.2). After tokenization is applied stemming. All stop words are replaced by terminal symbol *STOP*. ORs contain many numerical values, like clinical test results, vitals (Body Mass Index, Riva Roci - blood pressure), etc. Numerical values are replaced by terminal symbol *NUM*. The sentences have mainly telegraphic style, or the information is described as sequence of phrases separated by semicolon. We consider those phrases as *sentences*. Sentence splitting is applied to construct sequences of itemsets for each document. In this process all additional punctuation is removed. To separate the sentences is used negative number -1, and -2 is used to denote the end of the text. The last stage of the preprocessing is hashing, which purpose is to speed-up the process of frequent sequence mining. In the hashing phase each word is replaced by unique numerical ID.

For frequent sequence mining is used algorithm CM-SPAM [3], more efficient variation of SPAM algorithm [1], that is considered as one of the fastest algorithms for sequential mining. CM-SPAM is even faster than SPAM, but more important is that CM-SPAM is more efficient for low values of *minsup*. This is important, because in clinical text some cases are not so frequent, because the prevalence of the diseases is usually lower in comparison with other domains. The *minsup* values for clinical data are usually in the range [0.01,0.1].

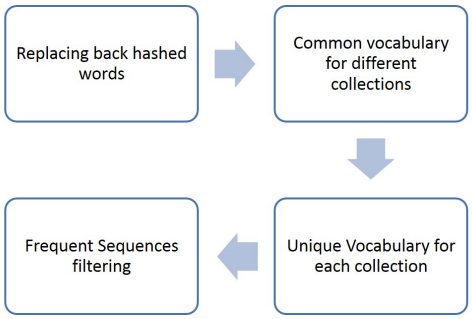

Fig. 3. Pipeline for postprocessing of the generated frequent sequences

The last step is the postprocessing phase (see Fig.3) that starts with replacing back the hashed words. Then we identify unique vocabulary for each collection:

$$WAU_i = WA_i - \bigcup_{j \neq i} WA_j - \bigcup_j WH_j$$
$$WHU_i = WH_i - \bigcup_{j \neq i} WH_j - \bigcup_j WA_j$$

Let $FA_1, FA_2, \ldots, FA_r$ and $FH_1, FH_2, \ldots, FH_r$are the frequent sequences generated on step 3. We need to filter all sequences that occur in any sequence of the other sets or a frequent sequence from other collection occur in them.

$$FFA_i = \{Z | Z \in FA_i \ \wedge \ \nexists j \neq i \ \ Y \in FA_j \vee Y \in FH_j$$
$$such\ that\ Y \subseteq Z \vee Z \subseteq Y \wedge \ \nexists X \in FA_i \ X \subseteq Z\}$$

$$FFH_i = \{Z | Z \in FH_i \ \wedge \ \nexists j \neq i \ \ Y \in FA_j \vee Y \in FH_j \\ such\ that\ Y \subseteq Z \vee Z \subseteq Y \wedge \ \nexists X \in FH_i \ X \subseteq Z\}$$

The so filtered frequent sequences sets together with unique words form the specific terminology and sub-language used by different specialist in patient disease history and status description.

6 Experiments and Results

For a cohort of 300 patients suffering from T2DM and COPD are extracted ORs for all their clinical visits in 3 year period (2012-2014) to different specialists: Otolaryngology (S14), Pulmology (S19), Endocrinology (S05), and General Practitioners (S00). After preprocessing of ORs in all collections are separately extracted Anamnesis and Status descriptions for each patient (Table 2 and Table 3).

The minsup value were set as relative minsup function of the ration between the number of patients and ORs. It is approximately 0.02% for the smallest set SA14, 0.03% for SA05 and SA19 and 0.1% for the largest set SA00. This is a rather small minsup value that will guarantee coverage even for more rare cases but with sufficient support. For Status subset the minsup value were set in similar range – 0.05% for SH14 and SH19, 0.08% for SH05 and 0.09% for SH00.

All subsets are processed with CM-SPAM for frequent sequences mining. In addition the algorithm dEclat [18] for frequent itemsets mining was applied. The frequent itemsets were filtered with similar method as frequent sequences (see Table 2 and Table 3). For experiments are used Java implementations of the algorithms from SPFM (Open-Source Data Mining Library) [6].

	SA00	SA05	SA19	SA14
ORs	11,345	798	532	156
Patients	294	195	70	173
Items / Vocabulary	4,337	1,767	1,527	447
minsup	0.1 (131)	0.03 (24)	0.02(11)	0.03(5)
Frequent Sequences	1,713	23,677	8,250	6,643
Filtered Sequences	1,358	23,327	7,932	6,527
Frequent Itemsets	80	1,815	477	200
Filtered Itemsets	37	1,732	396	178
Unique words	2,923	747	628	144

Table 2. Frequent sequences in Anamnesis section

[6] SPFM, http://www.philippe-fournier-viger.com/spmf/index.php?link=algorithms.php

	SH00	SH05	SH19	SH14
ORs	11,345	798	532	156
Patients	294	195	70	173
Items / Vocabulary	3,412	1,131	700	627
minsup	0.09 (1,022)	0.08 (64)	0.05 (27)	0.05 (8)
Frequent Sequences	107,267	27,949	26,341	345
Filtered Sequences	106,634	27,185	25,670	321
Frequent Itemsets	31,902	7,176	2,224	30
Filtered Itemsets	30,462	5,551	1,467	22
Unique words	2,422	391	195	346

Table 3. Frequent sequences in Status section

The datasets for Anamnesis are sparse, because they contain descriptions of different patient diseases history, complaints, and risk factors. Thus the diversity of explanations causes lower number of generated frequent sequences and higher number of unique vocabulary (see Fig.4 and Table 2). The unique vocabulary contain different complaints and many informal words for their explanation. Although the set SA00 is larger than the other sets for this set are generated lower number of frequent sequences. This set corresponds to the ORs written by general practitioners, who usually ob-serve larger set of diseases than other specialists. The set SA05 contains more consistent information about the T2DM complaints only.

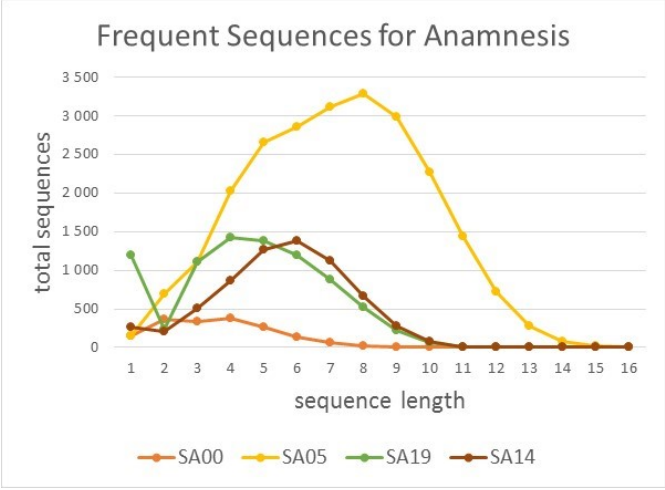

Fig. 4. Generated frequent sequences for Anamnesis section grouped by length

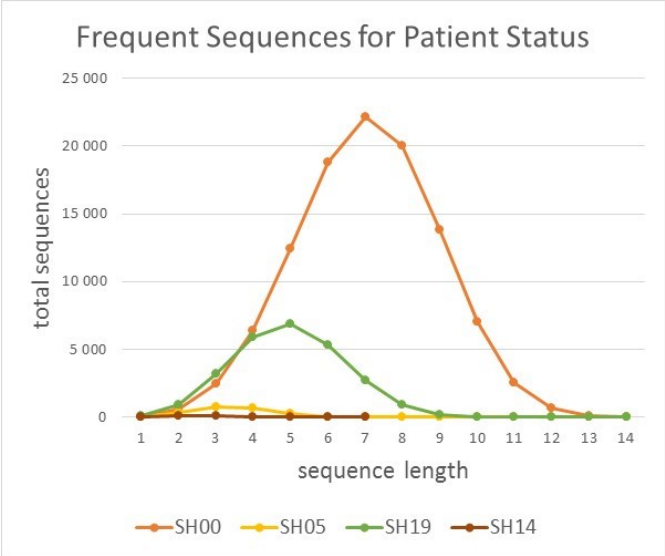

Fig. 5. Generated frequent sequences for Status section grouped by length

In contrast the datasets for Status are dense, because they contain predefined set of organs and systems status description. The Status explanation usually contains phrases rather than sentences. Each phrase describes single organ/system and its current condition. The similarity between Status explanations causes significant growth of the number of generated frequent sequences and lower number of unique vocabulary (see Fig.5 and Table 3). Although the higher number of the generated frequent sequences during the filtering process they shrink faster, because contain similar subsequences. The unique vocabulary contains specific terminology for some organs and systems that are in main focus and interest for the physician that makes the medical examination. The result set of context sequence contain only specific sub-language used from specialists in their area.

The extracted frequent sequences and frequent itemsets are used for multi class text classification. Experiments are provided by non-exhaustive cross-validation (5 iterations on sets in ratio 7:1 training to test). For comparison of the obtained results is used bag of words (BOW) method by applying frequent itemsests generated by dEclat algorithm.

The classification is based on unique vocabulary used for classes and on the filtered sequences and frequent itemsets from all classes that match the text. As golden standard in the evaluation are used specialty codes from ORs structured data.

Six types of experiments are performed. In the first task are used subsets for Anamnesis section for all four specialty classes 00, 05, 14 and 19. The evaluation results (Table 4) for F1 measure ($F1 = 2 * Precission * Recall/(Precission + Recall)$) show that context sequences method outperforms BOW method for all

classes, except class 19 for Anamnesis subsets. The evaluation for Status section classification is just the opposite (Table 6). BOW method shows better results than context sequences. The main reason is that Status section is written in telegraphic style with phrases rather than full sentences. Usually Status section contains sequence of attribute-value (A-V) pairs - anatomical organ/system and its status/condition.

General practitioners used in ORs terminology and phrases that can be found in ORs for all specialties. Thus the class 00 is not disjoint with classes 05, 14 and 19. Class 00 is one of the main reasons for misclassification. Another experiment was performed with "pure" classes – including only 05, 14, and 19 (Table 5). The F1-measure values show better performance in classification task for Anamnesis for all classes, in compassion with BOW method. For Status tast the results for both methods are comparable (Table 7).

	Context Sequences				BOW			
	SA00	SA05	SA14	SA19	SA00	SA05	SA14	SA19
Precision	0.9986	0.3674	0.8707	0.6130	0.9997	0.1872	0.7785	0.7804
Recall	0.8574	0.9848	0.8205	0.9568	0.6944	0.9975	0.7436	0.8684
F1	0.9226	0.5351	0.8449	0.7473	0.8195	0.3152	0.7607	0.8221

Table 4. Evaluation of rules for Anamnesis for S00, S05, S14 and S19

	Context Sequences			BOW		
	SA05	SA14	SA19	SA05	SA14	SA19
Precision	0.9581	1.0000	0.9714	0.8803	1.0000	0.9935
Recall	0.9873	0.8312	0.9751	0.9987	0.7436	0.8701
F1	0.9725	0.9078	0.9733	0.9358	0.8529	0.9277

Table 5. Evaluation of rules for Anamnesis for S05, S14 and S19

	Context Sequences				BOW			
	SH00	SH05	SH14	SH19	SH00	SH05	SH14	SH19
Precision	0.9990	0.5551	0.9560	0.2420	0.9750	0.6105	1.0000	0.8954
Recall	0.8108	0.9010	0.9744	0.9925	0.9653	0.7995	0.9487	0.6891
F1	0.8951	0.6870	0.9651	0.3891	0.9701	0.6923	0.9737	0.7788

Table 6. Evaluation of rules for Status for S00, S05, S14 and S19

	Context Sequences			BOW		
	SH05	SH14	SH19	SH05	SH14	SH19
Precision	0.9902	1.0000	0.8829	0.9251	1.0000	1.0000
Recall	0.9103	0.9744	0.9944	1.0000	0.9610	0.8910
F1	0.9486	0.9870	0.9353	0.9611	0.9801	0.9424

Table 7. Evaluation of rules for Status for S05, S14 and S19

	Context Sequences				BOW			
	S00	S05	S14	S19	S00	S05	S14	S19
Precision	0.9999	0.6958	0.9750	0.6251	0.9979	0.8356	0.9935	0.9618
Recall	0.9428	0.9975	1.0000	1.0000	0.9871	1.0000	0.9809	0.9097
F1	0.9705	0.8198	0.9873	0.7693	0.9924	0.9105	0.9872	0.9351

Table 8. Evaluation of rules for ORs for S00, S05, S14 and S19

	Context Sequences			BOW		
	S05	S14	S19	S05	S14	S19
Precision	1.0000	1.0000	0.9981	0.9645	1.0000	1.0000
Recall	0.9987	1.0000	1.0000	1.0000	0.9872	0.9492
F1	0.9994	1.0000	0.9991	0.9819	0.9935	0.9739

Table 9. Evaluation of rules for ORs for S05, S14 and S19

Finally the classification of both sections – Anamnesis and Status is used for classification of the outpatient record as a whole document. The evaluation results (Table 8) show that results for context sequences drop down and BOW method performance is better. After eliminating the noisy set S00 - the result (Table 9) for context sequences method significantly improve and outperform BOW method for all three classes 05, 14 and 19.

7 Conclusion and Further Work

The proposed data-driven method is based on data mining techniques for context sequences identification in clinical text depending on medical specialty of the doctor. The method is language independent and can be used for low resource

languages. The huge number of generated frequent sequences is reduces during the filtering process. The experimental results show that context sequences methods outperforms BOW method for sparse datasets in classification task.

Using "human-in-the-loop" [6] approach some further analyses of the significance for the domain of the generated frequent sequences and the misclassified documents will be beneficial. The space of clinical events is too complex. Thus "human-in-the-loop" can be applied also for subclustering task by using patient age, gender and demographic information. Reducing the dimensionality will help to determine different context sequences depending on the patient phenotype.

As further work can be mentioned also the task for context sequences similarities measuring. It can be used to identify synonyms and semantically close phrases.

Acknowledgments

This research is partially supported by the grant SpecialIZed Data MIning MethoDs Based on Semantic Attributes (IZIDA), funded by the Bulgarian National Science Fund in 2017–2019. The author acknowledges the support of Medical University - Sofia, the Bulgarian Ministry of Health and the Bulgarian National Health Insurance Fund.

References

1. Ayres, J., Flannick, J., Gehrke, J., Yiu, T.: Sequential pattern mining using a bitmap representation. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 429–435. ACM (2002)
2. Boytcheva, S., Angelova, G., Angelov, Z., Tcharaktchiev, D.: Integrating data analysis tools for better treatment of diabetic patients. CEUR Workshop Proceedings **2022**, 229–236 (2017)
3. Fournier-Viger, P., Gomariz, A., Campos, M., Thomas, R.: Fast vertical mining of sequential patterns using co-occurrence information. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 40–52. Springer (2014)
4. Ge, J., Xia, Y., Wang, J., Nadungodage, C.H., Prabhakar, S.: Sequential pattern mining in databases with temporal uncertainty. Knowledge and Information Systems **51**(3), 821–850 (Jun 2017). https://doi.org/10.1007/s10115-016-0977-1, https://doi.org/10.1007/s10115-016-0977-1
5. Gunning, D.: Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA), nd Web (2017)
6. Holzinger, A.: Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Informatics **3**(2), 119–131 (2016)
7. Huang, J., Huan, J., Tropsha, A., Dang, J., Zhang, H., Xiong, M.: Semantics-driven frequent data pattern mining on electronic health records for effective adverse drug event monitoring. In: Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on. pp. 608–611. IEEE (2013)
8. Jensen, K., Soguero-Ruiz, C., Mikalsen, K.O., Lindsetmo, R.O., Kouskoumvekaki, I., Girolami, M., Skrovseth, S.O., Augestad, K.M.: Analysis of free text in electronic health records for identification of cancer patient trajectories. Scientific reports **7**, 46226 (2017)

9. Jindal, R., Malhotra, R., Jain, A.: Techniques for text classification: Literature review and current trends. webology **12**(2), 1 (2015)
10. Krumholz, H.M.: Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Affairs **33**(7), 1163–1170 (2014)
11. Névéol, A., Dalianis, H., Velupillai, S., Savova, G., Zweigenbaum, P.: Clinical natural language processing in languages other than english: opportunities and challenges. Journal of biomedical semantics **9**(1), 12 (2018)
12. Patnaik, D., Butler, P., Ramakrishnan, N., Parida, L., Keller, B.J., Hanauer, D.A.: Experiences with mining temporal event sequences from electronic medical records: initial successes and some challenges. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 360–368. ACM (2011)
13. Plantevit, M., Charnois, T., Klema, J., Rigotti, C., Crémilleux, B.: Combining sequence and itemset mining to discover named entities in biomedical texts: a new type of pattern. International Journal of Data Mining, Modelling and Management **1**(2), 119–148 (2009)
14. Rabatel, J., Bringay, S., Poncelet, P.: Mining sequential patterns: a context-aware approach. In: Advances in Knowledge Discovery and Management, pp. 23–41. Springer (2013)
15. Wang, F., Lee, N., Hu, J., Sun, J., Ebadollahi, S.: Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 453–461. ACM (2012)
16. Wright, A.P., Wright, A.T., McCoy, A.B., Sittig, D.F.: The use of sequential pattern mining to predict next prescribed medications. Journal of biomedical informatics **53**, 73–80 (2015)
17. Yadav, P., Steinbach, M., Kumar, V., Simon, G.: Mining electronic health records (ehrs): A survey. ACM Computing Surveys (CSUR) **50**(6), 85 (2018)
18. Zaki, M.J., Gouda, K.: Fast vertical mining using diffsets. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 326–335. ACM (2003)
19. Ziembiński, R.Z.: Accuracy of generalized context patterns in the context based sequential patterns mining. Control and Cybernetics **40**, 585–603 (2011)