Mining Association Rules from Clinical Narratives

Svetla Boytcheva **Ivelina Nikolova** **Galia Angelova**
Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
25A Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
svetla.boytcheva@gmail.com {iva,galia}@lml.bas.bg

Abstract

We propose a method that processes raw informal medical texts (from health forums) and formal texts (outpatient records) in Bulgarian language in order to extract typical word co-occurrences in the form of association rules. When mining these rules we use some context information and small terminological lexicons to generalize the extracted frequent patterns. This allows to study informal expressions of medical terminology and to identify automatically typical descriptions of types of patient statuses. The paper presents association rules generated from 300,000 outpatient records and 1,425 forum postings and discusses their evaluation and usefulness. Employing this unsupervized data mining approach we hope to overcome the lack of linguistic resources that can support successful NLP analysis of clinical narratives in Bulgarian.

1 Introduction

Clinical narratives written by medical experts are a growing source of patient-related information that can be used in medical research and healthcare management. In addition, informal medical texts and conversations in social networks and popular TV broadcasts increase significantly. Public forums for medical consultations exist as well where doctors provide opinions and answer questions. In this case medical experts abandon the professional style of clinical writings (brief reports in cryptic form with medical terminology and specific words, abbreviations, acronyms, and phrases of their professional jargon). They switch to informal, conversational style (avoiding terminology, using casual words for popular explanations, with shorter and simpler sentences etc.). Differences between informal and formal communication are often addressed in translation where the objective is to reproduce language in all its variety. In (Lozano and Matamala, 2009) the authors analyse how the original English medical terminology, used to give a special atmosphere and to reproduce a real professional context in the television series E.R., was translated in colloquial language to reach a broad audience of Spanish native speakers in the Spanish dubbed version. The original English terms were manually categorized based on their degree of formality in order to differentiate between formal and informal medical language. Then rules for selecting translations in different settings were elaborated. This structuring enabled to approach systematically the evaluation of translation. Many mistakes have been found in the translation of medical terminology, hence lowering the realism of the dubbed version. The conclusion is that terminology misuse is rather often in informal communication.

Here we also consider formal and informal medical texts but as input data for mining Association Rules (AR) that explicate stable collocational patterns of terms and words. The overall objective is to find frequent patterns of lexical co-occurrences which, as we suppose, will help to structure patient descriptions. In general there are no standard templates how patients are to be examined, e.g. in diabetes there is no predefined structured questionnaire how to document the status of a diabetic patient in the form of indicators and their values. Earlier attempts to structure patient status descriptions took us months to study a large number of patient records and weeks of manual work to produce feature-value templates that were discussed with medical experts (Boytcheva et al., 2010). Now we believe that ARs will discover automatically attributes and their values as

Proceedings of Recent Advances in Natural Language Processing, pages 130–138,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_019

the latter are expressed by lexical units in the free text descriptions. Therefore it is important to consider patient-related texts in records produced by medical doctors as well as informal texts written by the patients themselves. In addition for low resource languages like Bulgarian, Data Mining provides better instruments to discover new knowledge than Text Mining via shallow analysis and information extraction techniques. We propose a method that processes raw formal clinical narratives (outpatient records, OR) and informal texts (postings in health forums) in Bulgarian language. Some context information is taken into consideration when mining the ARs, which is an original aspect of the proposal. Small terminological lexicons provide generalization of extracted frequent patterns. This approach allows to map informal expressions of medical terminology to the formal ones and to study in parallel both the professional and colloquial medical language.

The paper is structured as follows. Section 2 overviews related work. Section 3 presents the materials used and Section 4 - the methods. Section 5 details the experiments. Section 6 contains the conclusion and plans for future work.

2 Related Work

Almost no electronic resources with medical terminology exist for Bulgarian language (except for terms in standard medical nomenclatures) so we are interested in terminology extraction for low resource languages. For Polish, linguistic analysis and statistical methods identify automatically phrases that cover 84% of the occurring medical terms in over 1,200 discharge letters (Marciniak and Mykowiecka, 2014). At the top of the ranked list, only 4% out of 400 terms were incorrect. Another work deals with term extraction from sparse, ungrammatical and informally-written texts with domain-specific contents (Ittoo and Bouma, 2013). This paper focuses on rare (low frequency) terms, detects multi-word terms of arbitrarily lengths which are often disregarded by existing term extraction systems, and involves external resources (Wikipedia) to support domain-specific term extraction and assessment of accuracy. A rather high F1-measure is achieved (88% against a baseline of 77%) and successful extraction of terms regardless of their length.

Discovering frequent word sequences also provides useful hints about analysis of units in formal and informal texts. A method for extraction of all maximal frequent word sequences, which allows for gaps, is presented in (Ahonen-Myka, 1999). The algorithms in (Ahonen-Myka, 2002) include pruning of all stop words, which might be problematic in case of terminological expressions. No stemming is applied since inflexion endings might be meaningful, moreover their removal will combine sequences and in this case some low-frequency variations can exceed together the selected frequency threshold. In (Ahonen-Myka and Doucet, 2005) the discovery of maximal frequent word sequences is re-considered in the light of collocation discovery, where "collocation" is a recurrent, stable multiword expression without gaps. The authors assume that most existing methods for collocation discovery cannot be straightforwardly extended to find sequences with length more than 5 words, when applied to large corpora.

Text mining extracts essential information from texts while data mining (in particular ARs) discovers novel knowledge about the subject. ARs that are found in texts are used in various kind of document-processing applications, for instance:

(i) Text classiffication based on ARs: sentences of documents are viewed as basic text units (Haralambous and Lenca, 2014). This approach is enriched by linguistic knowledge (delivered by the Stanford dependency parser). Words are replaced by their hyperonyms in WordNet, to optimize the itemsets. In the training phase ARs are mined from sentences. At the classification stage, for each sentence s in a document d the system finds the most confident AR that can be applied to it (i.e., such that the itemset of the rule is entirely contained in the itemset of the sentence). An aggregation procedure classifies the document by taking class by class the sum of rule confidence and selecting the class with the highest sum. The evaluation was done on 7,000 texts of Reuters corpus. The experiments show that using dependency property (*nsubj*) to select a word is a better choice than the one provided by the *tf-idf*-based method. However, no alternative classification techniques are applied in parallel to the same text collection so no real conclusion can be drawn about the effectiveness of the suggested approach.

(ii) Elucidating domain concepts based on ARs: the paper (Yolcular, 2011) presents 12 ARs that can point to significant medical concepts mined in 600 otorhinolaryngology discharge notes written

in Turkish language. The *n*-gram method was used for discovering terms co-occurrences. A dataset of concept candidates has been generated for the validation step and then the Predictive Apriori algorithm for AR mining was applied to validate the candidate concepts.

(iii) Explication of relations among text units: (Sizov and Öztürk, 2012) present the SmoothApriori algorithm that finds association relations between sentences which may reveal a cause-effect type of relation or have a more implicit nature. SmoothApriori uses similarity between items; compared to a previously proposed SoftApriori algorithm, which also makes use of similarity, it is able to utilize similarity values directly rather than reducing them to binary values similar/not-similar. The evaluation was done on "Findings as to Risk" section of 208 Air Investigation Reports, published by the Transportation Board of Canada. Many top confidence rules, automatically generated by SmoothApriori, are interesting and make sense. Some rules connect consecutive sentences in the same text but the implicit discourse relation of causality is explicated. This application illustrates the potential of using ARs in various areas.

The results presented here integrate text and data mining ideas, extending further our previous developments (Boytcheva et al., 2017).

3 Materials

We work with two data sources: ORs submitted to the Bulgarian National Health Insurance Fund (NHIF) and content from the online medical portal *puls.bg*[1]. The ORs concern 10,000 diabetic patients and were produced by Endocrinologists (ESs) (set **S00**) and General Practitioners (GPs) (set **S05**) in 2012–2013. In total these are 330,666 records, semi-structured files with predefined XML-format. We use only two free text fields of these ORs: "*Anamnesis*" and "*Patient Status*". The corpus of informal medical texts includes postigns at *puls.bg* forum (set **SF**). We process all questions and comments from three subforums[2] amounting to 1,425 records.

Fig. 1 presents the distribution of words (items) in the sets after stemming and stop words removal.

We analyzed how the sources vocabularies are distributed in different categories using available lexicons. Table 1 presents the distribution of categories in subsets of 20% of the vocabulary in $S00$, 35% in $S05$ and 55% in SF.

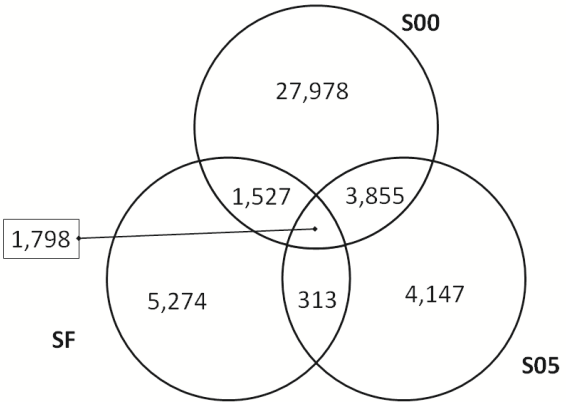

Figure 1: Corpus vocabulary

Set	$S00$	$S05$	SF
common words	10.7%	15.1%	36.6%
medical terms	1.2%	1.6%	1.2%
names of diseases	5.8%	10.9%	12.0%
treatment	0.4%	0.8%	0.7%
symptoms	0.8%	1.9%	1.7%
abbreviations	0.6%	1.4%	0.6%
not classified	79.6%	66.8%	44.2%

Table 1: Vocabulary distribution in training sets

The items of SF are mainly general lexica (36.6%) and names of diseases (12.2%). The most frequent units are stop words (43%) followed by numerical values (5%), some common words like have/has, year, value, month, examination, problem, result etc. signaling for the duration of the symptoms and medical terms signaling for the location of the finding - endocrine, hormon, thyroid gland.

The items of $S00$ are mainly common words and names of diseases. The most frequent units are stop words (19.84%), numerical values (7.54%), names of anatomic organs and systems as well as their conditions: breath, abdomen, liver, rhytmic, soft, RR (Riva Roci), vesicular, pulmo, limbs.

Similarly to $S00$, the items of $S05$ are mainly common words and names of diseases. Some of the most frequent units are stop words (17.12%), numerical values (12.60%), and indicators related to diabetes concepts like: sugar, treatment, activity, diabetes, rythmic, type, breath, BSP (blood sugar profile).

The analysis of corpus items (Table 2) shows that in the intersection $S4$ of the three training corpora, some 34% are general words from the top

[1] http://puls.bg

[2] Diabetes, Smoking cessation, Thyroid gland

Set	$S1$	$S2$	$S3$	$S4$
common words	39.5%	22.1%	34.8%	34.0%
medical terms	1.8%	2.2%	1.7%	1.7%
names of diseases	25.4%	18.1%	30.4%	32.3%
treatment	1.4%	1.3%	1.6%	1.7%
symptoms	4.3%	3.4%	5.5%	6.1%
abbreviations	1.4%	2.2%	1.8%	1.8%
not classified	24.2%	48.6%	22.3%	20.6%

Table 2: Common vocabulary in the training sets, where $S1 = S00 \cap SF, S2 = S00 \cap S05, S3 = S05 \cap SF, S4 = S00 \cap S05 \cap S0F$

10,000 most frequent words in Bulgarian, names of diseases (32.2%), symptoms (6.1%), abbreviations (1.8%) and other medical terms (1.7%).

The common items for SF and $S05$ represent terminology related to diabetes. This is not surprising because the forum topic is about diabetes and $S00$ contains ORs written by Endocrinology specialists. Examples of such items are: enlarged, weight, hemoglobin, insulin, examination, pain, control, glycated, treatments, blood, sugar, consisting, profile, diabetes etc.

The common items for SF and $S00$ are mainly terms related to medical examinations. This is also easy to explain since SF consists of informal texts and $S00$ contains clinical notes of General Practitioners. Examples of items common for SF and $S00$ are: treatment, procedure, drugs, complains, pain, examination, condition, consultation, normal, increased, control, changed, etc.

The common items for $S05$ and $S00$ are mainly terms concerning patient status - names of anatomic organs and systems. Examples of such items are: succusio renalis, palpated, non-painful, enlarged, ripple, pink, terminal, height, weight, limbs, pulmonary, good, breathe, family, edema, vesicular, complained, RR, liver, neck, rhythmically, peripheral, diabetes, control, heart, etc.

4 Text Analysis Methods

Our approach has three main phases: *preprocessing* which converts the text documents into itemsets, *processing* based on frequent pattern mining (FPM) techniques and elicitation of ARs, and *postprocessing* that filters, maps and generalizes rules by using context information and small lexicons (Fig. 2). The system processes input texts in unicode format and is language independent in principle (stemming and stopword filtering can be replaced by modules for another language).

4.1 Preprocessing

We have three text collections: *SF* - questions and comments in forum postings, *S05* and *S00* - the free texts of "*Anamnesis*" and "*Patient Status*" sections of ORs written by ESs and GPs correspondingly. Each text in *SF*, *S05*, and *S00* is turned to a sequence of word stems in their original order, using blank spaces and punctuation delimiters as tokenization separators. Stop words and numbers may be essential for some patterns so they are preserved and generalized - replaced by the constants STOP and NUM correspondingly. After this step the punctuation is eliminated because it is often erroneosly written or missing in both forums and in ORs.

Let *S* be one collection. The vocabulary used in all documents of *S* will be called *items* $W = \{w_1, w_2, ..., w_n\}$. For the collection S we extract the set of all different documents $P = \{p_1, p_2, ..., p_N\}$, where $p_i \subseteq W$. This set corresponds to transactions; the associated unique transaction identifiers (*tids*) shall be called ***pids*** (patient identifiers). Each patient interaction with a doctor (question or comment in *SF* or an anamnesis or patient status section of an OR in *S00* and *S05*) is viewed as a single document in *P*.

4.2 Processing

Our documents are written in different styles: the forum texts have quite informal syntax structure while the ORs are written in telegraphic style with phrases rather than full sentences. Usually the ORs list attribute-value (*A-V*) pairs - anatomic organs/systems and their status/condition:

$$A_1V_1, ..., A_nV_n | V_1A_1, ..., V_nA_n \ .$$

E.g.: Кор - ритмична нормофреквентна сърдечна дейност, Крайници- хипестезия от дистален тип, везикуларно дишане. (*Cardiovascular system - rhythmic norm frequent heartrate, limbs - hypoesthesia distal type, vesicular breath.*)

where A_1=*Cardiovascular system*, A_2=*limbs* and A_3=*breath* are attributes and their corresponding values are *V1=rhythmic norm frequent heartrate*, *V2=hypoesthesia distal type* following A_1 and A_2 and *V3=vesicular* preceeding A_3.

Attribute names contain phrases and abbreviations in Cyrillic and Latin. Values can be long descriptions in case of status complications.

The order of *A-V* pairs can vary and parts of the value descriptions can surround the attributes:

$$V_1...V_kAV_{k+1}...V_n.$$

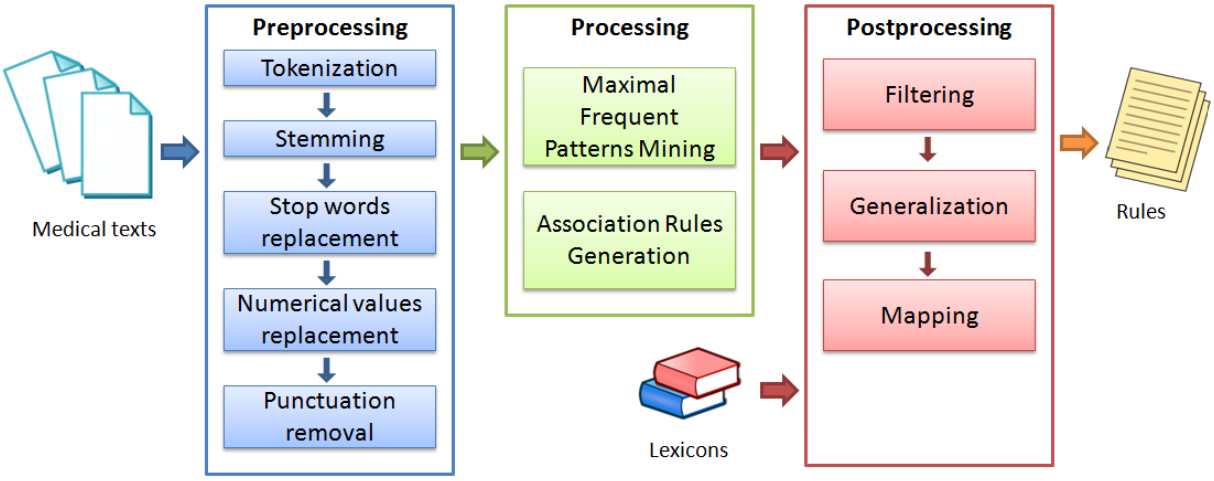

Figure 2: System Architecture

E.g.: Об.везикуларно дишане, средни и дребни хрипчета в белодробните основи... (*Ac.vesicular breathing, moderate and small wheezes in the lungs' bases...*)

where *A*=*breathing* and the surrounding words describe the status (attribute value).

It is also possible that some attributes share the same value:

$$A_1, A_2, ..., A_n V | V\, A_1, A_2, ..., A_n.$$

E.g.: хепар ет лиен неувеличени, Глава и шия - б.о., Кожа и видими лигавици бледорозови, ... (liver and spleen not enlarged, Head and neck - without peculiarities, Skin and visible mucous membranes pale pink, ...)

At the same time the informal texts in the health forum contain descriptions of patient status in a rather sparse form with much richer language and wider context around the triggering terms.

E.g.: Здравейте, моля ви кажете какво да правя, сърцето ми прескача от над 3 часа и то доста често... (*Hello, please tell me what to do, my heart skips over three hours and quite often...*)

Thus, when searching for frequent patterns, we consider a window of more than 10-12 words around each attribute. The rich terminology and flexible syntax structure hinder the application of traditional methods for extraction of collocations with gaps. These approaches would rather find the OR clishe phrases as collocations with highest frequency, moreover many *A*-*V* pairs would be erroneously considered as *n*-grams.

E.g.: **Positive examples**:
(forum) електронна цигара (*electronic cigar*)
(forum) щитовидна жлеза (*thyroid gland*)
(ORs) общо състояние (*general condition*)
(ORs) щитовидна жлеза (*thyroid gland*)
Negative examples:
(forum) имам Хашимото (*I have Hashimoto*)
(ORs) диабет тип (*diabetes type*)
(ORs) мек неболезнен (*soft non-painful*)

Therefore we treat documents as bag of words rather than sequences, they are transformed to itemsets with single word occurrences only.

Given a set of pids S, support of an itemset I is the number of pids in S that contain I. We denote it as $supp(I)$. We define a threshold called *minsup* (minimum support). A frequent itemset (FI) I is one with at least minimum support count, i.e. $supp(I) \geq minsup$. The task of FPM of S is to find all possible frequent itemsets in S.

Most FPM algorithms generate all possible frequent patterns (FPs). The search space grows exponentially with the size of W. Summarized information for data relations can be extracted as maximal frequent itemsets (MFI). The condensed information not only accelerates the process, reducing redundancy, but also decreases significantly the number of frequent patterns for post-analysis.

An implication in the form $I \Rightarrow J$ is called *association rule*, where $I \subset W, J \subset W, I \cap J = \emptyset$. I is called antecedent or ancestor and J is called consequent. Support of a rule is the number of pids in S that contain $I \cup J$, i.e.

$$sup(I \Rightarrow J) = sup(I \cup J) = P(I \cup J).$$

If *C%* of the documents in S that contain *I*, contain also *J*, then the association rule $I \Rightarrow J$ holds with *confidence* C in S, i.e. this is the condition probability

$$conf(I \Rightarrow J) = P(J|I) = \frac{sup(I \cup J)}{sup(I)}.$$

The task of ARs mining in collection S is to generate all ARs with confidence above the user defined confidence (*minconf*) and support above user defined support (*minsup*). Rules that satisfy both a minsup and minconf are called *strong*.

However, even for reasonable values of minconf and minsup, big datasets yield huge sets of strong ARs. Thus we can use an additional filter called *lift* that is defined as the ratio of the confidence of the rule and the confidence of its consequent.

$$lift(I \Rightarrow J) = \frac{P(I \cup J)}{P(I)P(J)}.$$

The lift represents the strenght of the relation between the consequent and its antecedent. A lift value < 1 indicates independence between them. If the lift value is > 1, this indicates that the antecedent and consequent appear together more often than expected, i.e. are correlated. Such rules are potentially usefull for predicting the consequent in new sets.

For ARs generation we use algorithms for mining all ARs with the lift measure in a transaction database (Agrawal and Srikant, 1994) with implementation at SPMF[3]. In the experiments we applied the FPmax algorithm (Grahne and Zhu, 2003) for MFI and All Association Rulse with FP-Growth with lift (Han et al., 2004).

4.3 Postprocessing

In order to find certain correlation among rare items, the minimal support needs to be set rather low. This causes generation of a huge amount of ARs and most of them are redundant. Adapting methods of Ashrafi et al. (2007), we apply some techniques for redundant ARs removal. In addition we select only those ARs that fulfill the requirements to have support, confidence and lift above predefined thresholds minsup and minconf. The lift value is very important because it gives additional information about the usefulness of the generated rules. Thus we filter only ARs with lift > 1. This is a necessary condition but not a sufficient one. For instance we obtain ARs like:

везикулар кожа => *STOP*

where STOP is a marker for a stop word, "везикулар" means "vesicular" (concerning breath) and "кожа" means "skin". This rule has high support (about 8%), confidence (0.999) and although its lift (1.429) is quite high, obviously it is useless. Thus we add some additional constraints, like removing all ARs with consequent that contains only the $STOP$ constant.

At the next step we perform AR generalization, based on small lexicons. We use some terms as seeds and rule based prediction about the features of other words that appear in similar rules at the same position. Initially generalization is applied for symptoms/conditions and complains:

- For association rules $R_1 : I \cup X \Rightarrow J$ and $R_2 : I \cup Y \Rightarrow J$ in case X and Y are symptoms/conditions for the same anatomic organ/systems C, we replace them by the marker "$STATUS(C)$" and define a more general rule $R_C : I \cup STATUS(C) \Rightarrow J$.
- For association rules $R_1 : I \Rightarrow J \cup X$ and $R_2 : I \Rightarrow J \cup Y$ in case X and Y are symptoms/conditions for the same anatomic organ/systems G, we replace them by the marker "$STATUS(G)$" and define a more general rule $R_G : I \Rightarrow J \cup STATUS(G)$.

E.g.: R_1: лигавици => розови (*Oral mucosa => pink*)
R_2: лигавици => бледи (*Oral mucosa => pale*)
R_G: лигавици => STATUS(G)
(*Oral mucosa => STATUS(G)*)
G=лигавици (*Oral mucosa*)

The variety of all status conditions is huge but some complications are rare and it is unlikely to have them all included in ARs. The main generalization advantage is that more general rules will help to predict/recognize some status conditions in the text that are not included in our original lexicon. The main disadvantage is that not all markers $STATUS(i)$ are equivalent, i.e. the status descriptions for different anatomic systems and organs can differ. Thus we can not apply further generalization of already generalized rules.

Finally the ARs generated for different collections are mapped to each other in order to study the specifics of the extracted collocations in formal and informal medical language. As result we are able to enrich the possible contexts of the medical terminology occurring in both types of text (Figure 3). Thus we can define word embeddings for some terms included in ARs for all three sets. For some item $w_i \in W$ and collection S we define its context for all ARs in S such that: $C_S(w_i) = \{I | I \Rightarrow J \cup w_i\}$. In particular when $I \Rightarrow J \cup w_i$ holds in S then it also holds that $I \Rightarrow w_i$. Thus $C(w_i) = \bigcup C_S(w_i)$ is the observed context of item w_i for all collections. The observation of the terminology context will help for further study of its nature.

5 Experiments and Findings

The experiments were performed on the collections SF, $S00$ and $S05$. SF consists of informal

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

Set	$S05.1$	$S05.2$	$S05.3$	$S00.1$	$S00.2$	$S00.3$	$SF.1$	$SF.2$	$SF.3$
pids	48,129	48,129	48,129	215,326	215,326	215,326	1,141	1,141	1,141
FI	465,454	359,032	342,118	421,238	386,132	344,215	2,161	1,977	1,977
MFI	759	500	388	1,659	1,619	1,605	453	455	439
AR	4,985,268	3,084,677	3,482,135	1,596,953	1,455,634	1,345,088	202	147	176
minsup	0.05	0.05	0.05	0.04	0.04	0.04	0.02	0.02	0.02

Table 3: Generated association rules with minconf=1.0 and minlift=1.1 for 9 training sets

Set	$S05.1$	$S05.2$	$S05.3$	$S00.1$	$S00.2$	$S00.3$	$SF.1$	$SF.2$	$SF.3$
pids	12,032	12,032	12,032	53,831	53,831	53,831	284	285	285
all	99.98%	99.99%	99.98%	99.99%	99.99%	99.99%	95.43%	96.01%	97.34%
max 2,000	99.96%	100.00%	99.95%	100.00%	100.00%	99.99%	95.43%	96.01%	97.34%

Table 4: Evaluation of the generated association rules for 9 test sets

text and is rather small. $S00$ and $S05$ contain sections of ORs. $S00$ corresponds to ORs produced by GPs and is significantly larger than the other two collections. It presents more sparse information than $S05$ which contains more focused information about the specific domain of diabetes. For each of SF, $S00$ and $S05$ three experiments are provided by non-exhaustive cross-validation (3 iterations on sets in ratio 4:1 training to test). The sets are denoted by their original set name and the number of experiment, e.g. for the set $S05$ there are three training sets $S05.1$, $S05.2$, and $S05.3$.

Table 3 presents numbers of constructions found in the FPM experiments: pids, frequent itemsets, MFI and ARs. In ARs generation for all collections we used minconf=1.0 but different minsup depending on the sets' size. We considered only ARs with lift > 1.1.

Filter the rules with lift <1.1 and consequent STOP automatically reduced approx. 50% of the ARs for $S05$ and $S00$, where language is more formal with limited vocabulary and strong support. The same constraints lead to about 40% reduction of the ARs in SF which has rich vocabulary and small support. The distribution of the lift measure values in these two types of sets - formal vs informal was also quite different. Lift variation in $S05$ and $S00$ is much smaller than in SF.

Table 4 presents the evaluation of the generated ARs. Two types of tests are performed: with all generated ARs and for the top 2,000 ARs according to their antecedents' cardinality. High precision in consequent prediction is seen for all sets.

Some ARs bring new knowledge about patient status description:

корем ==> мек неболезн
(*abdomen => soft non-painful*)

This rule has not too high support (about 5%) but its lift is higher than 1.1. It is a reasonable rule which connects lexical units describing an anatomic organ and its typical status.

The following AR has also very high lift - 12.21 and represents a set of attributes and values describing the condition of a diabetic patient. All of them are terms or typical phrasal expressions in the domain of diabetes.

тургор видим шия ссс запаз общо състояни ==>
розов кожа
(*turgour visible neck ccc preserved general condition => pink skin*)

In the following OR excerpt, items from the AR antecedent are highlighted in blue and the predicted consequent items are highlighted in pink:

Запазено общо състояние. Глава - склери чисти. Видими лигавици - розови. Кожа - розова , норм. тургор. Шия. -щит.жлеза- увеличена 1А. Не се палпират увеличени лимфни възли. Дихат. с-ма-чисто везикуларно дишане. ССС - ритмична нормофреквентна сърдечна дейност.
(*Preserved general condition. Head - clear ciliary body. Visible tissue - pink. Skin - pink , normal turgor. Neck -thyroid gland- enlarged 1A. Lymph nodes do not palpate enlarged. Respiratory system-clear vesicular breathing. Circulatory system - normal heart rhythm.*)

The support sets of the ARs present patients with different profiles. In general several ARs can represent the conditions of different anatomic organs and systems of one patient and thus the initial profiling groups can be partitioned into profiling subgroups. Thus patients can be clustered in groups with similar health condition. A more detailed analysis of patients that belong to several profile groups is a task for further investigation.

Although generalization did not decrease significantly the total number of ARs, the new general

rules help to process unseen status descriptions in other collections of medical documents.

ARs mapping. By collapsing equal ancestors of the ARs generated from the respective data sets we observe regularities which describe the medical language in the ORs on the one hand and the informal language from the medical forum on the other hand. Figure 3 presents ARs whose ancestor contain "thyroid gland". The larger ellipses denote the ancestors of the rules and the smaller ones linked through arrows are the consequents. $SFrule1$ and $SFrule2$ (in gray) are result of processing the SF dataset; $S05ruleset1$ and $S05ruleset2$ come from the dataset $S05$. Each of the rule sets represents several rules with equal ancestor.

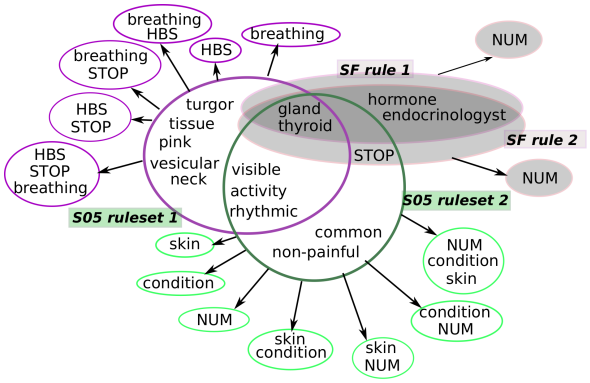

Figure 3: Mapping of ARs from forum data and ORs from Endocrinology professionals.

Among the extracted ARs from SF only 21 contain the terms "thyroid" and "gland" either in the ancestor or in the consequent whereas in the ARs resulting from $S05$ there are 3,985 ARs containing both terms. This shows that the informal language of non-specialists, although in medical topics, is rather difficult to formalize. The patterns in SF are short with up to 5 tokens in the ancestor vs. up to 14 tokens in the ARs coming from $S05$. Only 3 rules in SF have equal ancestors, in contrary the ancestors often can be collapsed in $S05$ and the same ancestor can have up to 62 different consequents. The ARs coming from $S05$ are also much longer, they describe related symptoms and/or diseases. As shown on Figure 3 the various consequents may be numerous but their vocabulary is not so rich. In $S05ruleset1$ there are only 3 tokens in the union of consequent vocabularies. Finally we note that the circles of mappings, as illustrated in Fig. 3, also suggest different profiles of patients experiencing eventual changes of thyroid.

6 Conclusion and Further Work

By fusing information from ORs written in a formal language with informal medical forum texts and applying frequent pattern matching techniques, we manage to extract and generalize ARs describing the context of medical terminology in the domain of diabetes. The obtained patterns have the means of stable term subsets which occur in the documents with a window longer than the usually utilized ones. In this sense our approach delivers higher benefits for generating resources that describe the terminology in the Diabetes domain in comparison to the word frequency based techniques such as *tf*, *tf-idf*. Moreover, the FP matching techniques have the advantage of being unsupervised thus they do not require any external knowledge or annotated data and can successfully deal with big data.

We generalize ARs for the common terms in three sets: terms related to diabetes generated by processing SF and $S05$ as well as terms related to medical examinations obtained by processing of SF and $S00$. We generalize also rules describing the context of terms describing the patient status, output from the processing of $S05$ and $S00$. These resources shall be further employed for automatic analysis of formal and informal medical records, symptoms and condition recognition.

Acknowledgments

The research presented here 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, and the project DFNP-100/04.05.2016 "Automatic analysis of clinical text in Bulgarian for discovery of correlations in the Diabetic Registry" funded by the Bulgarian Academy of Sciences in 2016-2017. The team acknowledges also the support of Medical University – Sofia, the Bulgarian Ministry of Health and the Bulgarian National Health Insurance Fund.

References

Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*.

Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, VLDB '94, pages 487–499. http://dl.acm.org/citation.cfm?id=645920.672836.

Helena Ahonen-Myka. 1999. Finding all maximal frequent sequences in text. In *Mladenic and Grobelnik (Eds.), Proc. 16th Int. Conf. on ML ICML-99, Workshop on ML in Text Data Analysis, J. Stefan Institute, Ljubljana*. pages 11–17.

Helena Ahonen-Myka. 2002. Discovery of frequent word sequences in text. In *Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery*. Springer-Verlag, London, UK, UK, pages 180–189. http://dl.acm.org/citation.cfm?id=647915.738872.

Helena Ahonen-Myka and Antoine Doucet. 2005. Data Mining Meets Collocations Discovery. In *Inquiries into Words, Constraints and Contexts, Festschrift in the Honour of Kimmo Koskenniemi*, CSLI Publications, Center for the Study of Language and Information, University of Stanford, pages 194–203. https://hal.archives-ouvertes.fr/hal-00324775.

Mafruz Zaman Ashrafi, David Taniar, and Kate Smith. 2007. Redundant association rules reduction techniques. *Int. J. Bus. Intell. Data Min.* 2(1):29–63. https://doi.org/10.1504/IJBIDM.2007.012945.

Svetla Boytcheva, Galia Angelova, Zhivko Angelov, and Dimitar Tcharaktchiev. 2017. Mining comorbidity patterns using retrospective analysis of big collection of outpatient records. *Health Information Science and Systems* (to appear). Springer Int. Publishing. https://link.springer.com/journal/13755.

Svetla Boytcheva, Ivelina Nikolova, Elena Paskaleva, Galia Angelova, Dimitar Tcharaktchiev, and Nadia Dimitrova. 2010. Obtaining status descriptions via automatic analysis of hospital patient records. *Informatica* 34(3):269–278. http://www.informatica.si/index.php/informatica/article/view/301/300.

Gosta Grahne and Jianfei Zhu. 2003. High performance mining of maximal frequent itemsets. In *6th Int. Workshop on High Performance Data Mining*. pages 135–143.

Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. 2004. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery* 8(1):53–87.

Yannis Haralambous and Philippe Lenca. 2014. Text classification using association rules, dependency pruning and hyperonymization. *Proc. of DMNLP, Workshop at ECML/PKDD, Nancy, France, CEUR Workshop Proceedings 1202, pp. 65-80* .

Ashwin Ittoo and Gosse Bouma. 2013. Term extraction from sparse, ungrammatical domain-specific documents. *Expert Systems with Applications* 40(7):2530 – 2540. https://doi.org/https://doi.org/10.1016/j.eswa.2012.10.067.

Dolores Lozano and Anna Matamala. 2009. The translation of medical terminology in tv fiction series: the spanish dubbing of e.r. *Vigo International Journal of Applied Linguistics* 6:73–87.

Małgorzata Marciniak and Agnieszka Mykowiecka. 2014. Terminology extraction from medical texts in polish. *Journal of Biomedical Semantics* 5(1):24. https://doi.org/10.1186/2041-1480-5-24.

Gleb Sizov and Pinar Öztürk. 2012. Mining of association relations in text. *Proc. Norsk informatikkonferanse, 2012, pp. 37-48, http://www.nik.no/2012/1-4-sizov12MiningOfAssociationRelationsInText.pdf* .

Basak Oguz Yolcular. 2011. Concepts extraction from discharge notes using association rule mining. *World Academy of Science, Engineering and Technology* (59):1031.