VERSITA

BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **12**, No 4

Sofia • 2012

Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2012-0030

Structured Information Extraction from Medical Texts in Bulgarian

Svetla Boytcheva

American University in Bulgaria, 1 Georgi Izmirliev Sq., Blagoevgrad 2700
Email: sboytcheva@aubg.bg

Abstract: *This paper presents an approach for Information Extraction (IE) from Patient Records (PRs) in Bulgarian. The specific terminology and lack of resources in electronic format are some of the obstacles that make the task of current patient status data extraction in a structured format quite challenging. The usage of N-grams, collocations and words' distances allows us to cope with this problem and to extract automatically the attribute-value pairs with relatively high precision.*

Keywords: *Artificial Intelligence, linguistic modeling, health informatics.*

1. Introduction

Patient Records (PRs) are the basic source of patient-related data, keeping all important medical information about each patient and providing access to the complete patient history. Usually this information is described only in the text and not presented in a structured format into the hospital information system, which does not allow to be automatically processed and to derive more complicated relationships between therapy condition, diagnoses and complaints. PRs status description contains a description of a local, somatic and specialized patient status. The main goal of our research is to extract patients' status data in a structured format (Attribute-Value). The "attributes" are anatomical organs, major anatomical system, their characteristics and physician examinations performed during the

admission. The "values" describe their actual condition for the patient. Thus, the structured presentation of the patient status can be presented as "attribute-value" tuples.

For detection of attributes and their values a statistical approach is used, which draws "useful" phrases by examining the frequency distribution of *N*-programs (sequences of *N* number of words), words collocations and words' distances in the text.

This paper is organized as follows: Section 2 describes the specifics of the PRs in Bulgaria and the data used for processing, Section 3 introduces *N*-grams and presents the used methods in more details, Section 4 reports results, discusses evaluation and related work, Section 5 contains a conclusion and sketches the further work.

2. Materials

In Bulgaria the discharge letter structure is mandatory for all hospitals (it is published in the Official State Gazette, as Article 190 (3) of the legal Agreement between the National Health Insurance Fund and the Bulgarian Medical and Dental Associations) [1]: personal details; diagnoses; anamnesis (personal medical history), including current complains, past diseases, family medical history, allergies, risk factors; patient status, including results from physical examination; laboratory and other tests findings; medical examiners comments; debate; treatment and recommendations.

The input texts in our experiment are free-text sections of discharge letters from Patient status section of PRs. The average number of sentences in the status section is 19.918, the minimal number is 8 sentences and the maximal number is 37 sentences. The training corpus contains 1300 PRs and the test corpus contains 6200 PRs with anonymised discharge letters provided by USHATE (University Specialized Hospital for Active Treatment of Endocrinology), Medical University, Sofia.

The various status descriptions present a number of key attributes, but there are attributes that are described only in cases where there are complications in the body. Examples for some common attributes are: gender, height, weight, bmi, skin, musculoskeletal system, limbs, and etc. In our corpus there are four types of the attributes and their values presentation [2]:

- **General** – by giving some default value, e.g., *без патологични промени, без особености* (*without pathological changes, without specifics*), or *със запазена/нормална характеристика* (*with preserved/present/normal characteristics*), etc.
- **Explicit** – the PR text contains particular specific values. The characteristic name might be missing since the attribute is sufficient to recognise the feature: e.g. "*preserved peripheral pulsations*" instead of "*preserved pulsations of the peripheral arteries*". The attributes are described by a variety of expressions, e.g., for the *"volume of the thyroid gland"* the value *"normal"* can be represented as *"not enlarged, not palpable enlarged, not palpable"*.

- **Partial** – The text contains descriptions about the organ parts, not about the main anatomical organ. For instance, the limbs status can be expressed like, e.g., *"atrophic changes of the legs skin with pretibial oedema"*.
- **By diagnosis** – sometimes a diagnosis is given instead of organ description, e.g., "*onychomycosis, tinea pedis*".

The main problem is that our corpus has an open vocabulary. Thus many "unknown" words can occur in the test corpus, but not to be presented in the training corpus. On the other hand, many rare attributes can be eliminated in the preprocessing phase due to low frequency.

This makes the task of automatically extracting pairs "attribute-value" quite complex without ontology of anatomical organs. The description in the status contains many terms in Latin, which further impedes the solution of the problem. The PRs contain mixed terminology both in Bulgarian and Latin Language and usage of Latin medical terms transcribed with Cyrillic letters. There are also many abbreviations both in Latin and Bulgarian. Further specific problems are due to the inflexional Bulgarian morphology; the terms occur in the text with a variety of word forms which is typical for the highly-inflexional Bulgarian language. The other obstacle is the lack of available resources in electronic format. Thus the task of extraction of the current patient status data in a structured format is quite challenging. The only advantage of PRs in Bulgaria is that the text is presented in a structured format using standard sections. This allows to split PRs into sections with high precision and to identify patient's status description.

3. Methods

There are several unsupervised and supervised approaches recently used for "attributes-value" pairs and other relation tuples extraction, such as: Maximum Entropy classifiers [3], Classifiers based on supervised methods [4], linguistics pattern-based relation extractor [5], semi-supervised relation extraction [6, 7]. *N*-grams approach is successfully used for "attributes-value" extraction in several applications, such as extracting information about geographic objects from Wikipedia [8], descriptions of products in Web pages [9, 10] and reviews [11].

N-grams can be used both [12] in symbolic and in word sequences methods. We use a statistical approach for extraction of useful phrases, based on the frequency distribution of *N*-grams (single words (unigram), word pairs (bigram), word triples (trigram) and word quadruples (quadrigram)).

Some examples for *N*-grams for attributes in our domain are:

- **single words (unigram)** – *ръст, тегло, итм, тургур, еластичност, глава, език, шия, слезка, крайници, корем (height, weight, BMI, turgor, elasticity, head, tongue, neck, spleen, legs, abdomen)*;
- **word pairs (bigram)** – *видими лигавици, очни ябълки, видима възраст, щитовидна жлеза, черен дроб, сукусио реналис (visible mucous membranes, eyeballs, apparent age, thyroid, liver, sucusio renalis)*;
- **word triples (trigram)** – *костно мускулна система, сърдечно-съдова система (musculoskeletal system, cardiovascular system).*

Our pipeline method performs the following steps (Fig. 1) over the 1300 PRs training corpus as input: (1) PRs sections splitting and identification of the Patient status section; (2) Collection of all status data from PRs; (3) Words extraction from all status data – set $S = \{w_1, w_2, \ldots, w_n\}$; (4) Filtering all numerical data from S – set Num and finding $S1 = S - Num$; (5) Filtering conjuncts and abbreviations – Set CA and finding $S2 = S1 - CA$ we don't lose precision, because we are interesting in "attributes" that are actually medical terms; (6) Filtering words with low frequency (rare words) – Set RW and finding $S3 = S2 - RW$; (7) Finding pairs (bigrams, 2-grams) – Set P; (8) Finding triples (trigrams, 3-grams) – Set T; (9) Filtering top *N* candidates from Set T – set T1; (10) Filtering top *N* candidates from Set P – set P1; (11) Filtering top *N* candidates from set S3 – set S4 (unigrams); (12) Selection of top *N* candidates for Attributes from P, T and S4 – Set A.

Fig. 1. *N*-grams filtering process

For Steps 9-11 28 special filtering rules are used and some general filtering rules, based on the frequency bellow some thresholds defined in advance. They filter the first trigrams candidates and during this process as a side effect also sets P and S3 are reduced, the latter is also changed during the pairs set filtering process, because all these three sets are strongly interrelated. Before discussing some filtering rules let us introduce some notations that can be used in the further explanations.

We will denote the word sequences either by $w_1 w_2 \ldots w_n$ or w_1^n. Notation where $p(w_i)$ represents the probability (frequency of occurrence) of the word w_i in our corpus (set S) and $p(w_i \mid w_1 w_2 \ldots w_n)$ is used for the probability of the

occurrence of the word w_i in case word sequence $w_1 w_2 \dots w_n$ is already available in the text, i.e., the probability the sequence $w_1 w_2 \dots w_n$ to be followed by the word w_i. To calculate $p(w_1 w_2 \dots w_n)$ the chain rule of probability is used:

$$p(w_1 w_2 \dots w_n) = p(w_1)p(w_2 \mid w_1)p(w_3 \mid w_1 w_2)\dots p(w_n \mid w_1^{n-1}) = \prod_{k=1}^{n} p(w_k \mid w_1^k).$$

Some filtering rules used in steps 9-11 are shown in (1)-(5):

(1) $$\text{if } p(w_1) > p(w_2 \mid w_1) \quad \text{then P} - \{w_1 w_2\},$$

(2) $$\text{if } p(w_2 \mid w_1) > p(w_3 \mid w_2) \text{ then P} - \{w_2 w_3\},$$

(3) $$\text{if } p(w_1) \approx p(w_1 w_2) \text{ then S3} - \{w_1\},$$

(4) $$\text{if } p(w_2 \mid w_1) > p(w_3 \mid w_1 w_2) \,\&\, p(w_2 \mid w_1) > p(w_3 \mid w_2) \text{ then } T - \{w_1 w_2 w_3\} \,\&\, \text{P} - \{w_2 w_3\},$$

(5) $$\text{if } p(w_2 \mid w_1) \approx p(w_3 \mid w_1 w_2) \text{ then P} - \{w_1 w_2\} \,\&\, \text{P} - \{w_2 w_3\},$$

where rule (1) is used for selecting unigrams "attribute" candidates, but it causes a reduction of the set P, rule (2) is used for bigrams selection – $w_1 w_2$ is the most stable pair compared to $w_2 w_3$ and it can be further used for P_1 candidates. The next rule (3) is used for selecting bigram "attribute" candidates, but it causes unigrams set S3 reduction, because $w_1 w_2$ is a stable pair in the corpus. Rules (4) and (5) are used for both trigrams and bigrams candidates filtering.

"Attribute" candidates are common words (with high frequency) for most of the patient status sections, thus they are mainly presented in the set S3. For some complications and disorders additional detailed explanations are added in the status data, containing rare "attributes" and currently they are filtered as rear words into the set RW. "Values" candidates can differ for different patients, so they should be mainly presented into the sets for Rare Words (RW) and NUMerical values (Num). We process the data for patients from a specialized hospital for endocrine disorders treatment and many patients have common symptoms and conditions thus some "values" are present with high frequency and currently they are filtered into the set S3.

In order to cope with the problem that sets RW and S3 can contain both attributes and values, we apply some additional methods for "attribute-value" tuples extraction.

"Values" selection procedure initially collects all numerical value from Num set, because they are specific for most of the patients and describe their current status. In the patient status description usually the "values" are surrounded (proceeded or followed) by attributes to which they correspond. Although they have different meaning, some numerical values can be used for several attributes and present data with different measures. For instance, 180 can be used both for height

in cm and for systolic blood pressure 180/100 mmHg. Thus, initially setting the "values" positions into the text and statistically finding their collocations (positions into the sentence) in the PRs sentences, we can obtain patterns for further identification of "attributes" potential collocations. Additional information usage of immediately following metrics after the numerical data helps to improve "attributes" identification. Due to the small number of used metrics into the corpus, they are manually added to the rules.

To cope with "attributes" with low frequency in our corpus and for the most common "values", patterns for "attribute-value" identification are automatically generated. The method for patterns generation is based on the word distances within the sentence. We assume that the information for some "attribute" and its corresponding value is described in the same sentence. For this phase only words from the set S1 are used (without numerical). From the PRs corpus we construct the set $\mathrm{C}=\{s_1, s_2, \ldots, s_m\}$ containing all sentences w_i from the patient status description sections. For each two words $w_i \in \mathrm{S1}, w_j \in \mathrm{S1}$ we find the set $\mathrm{C}' \subseteq \mathrm{C}$, such that each sentence $s_k \in \mathrm{C}'$ includes both of the selected words $w_i, w_j \in \mathrm{S1}$. We calculate the distance between these words: $d_{s_k}(w_i, w_j)$ for $\forall s_k \in \mathrm{C}'$. In case all calculated distances are the same, we denote them by $d(w_i, w_j)$ and construct a set of pattern $\mathrm{ptrn}(w_i, w_j)$. In case $w_i, w_j \in \mathrm{S1}$ are consecutive, then $d(w_i, w_j)=1$. Further we apply an aggregation procedure over the generated templates for pairs of words using the rules. For instance, for patterns $\mathrm{ptrn}(w_1, w_2)$, $\mathrm{ptrn}(w_2, w_3)$, $\mathrm{ptrn}(w_1, w_3)$ we can generate the pattern $\mathrm{ptrn}(w_1, w_2, w_3)$ only if the statement (6) is valid,

$$d(w_1, w_2) + d(w_2, w_3) = d(w_1, w_3). \tag{6}$$

In general for two patterns $\mathrm{ptrn}(v_1, v_2, \ldots, v_k)$ and $\mathrm{ptrn}(u_1, u_2, \ldots, u_l)$ where all $v_i, u_j \in \mathrm{S1}$, we order the words $v_1, v_2, \ldots, v_k$, $u_1, u_2, \ldots, u_l$ according to the distances $d(v_i, v_j)$, $d(v_i, u_j), d(u_i, u_j), d(u_i, v_j)$ between them into $w_1, w_2, \ldots, w_{k+l}$. We can generate the pattern $\mathrm{ptrn}(w_1, w_2, \ldots, w_{k+l})$ only if the statement (6) is valid for all w_i, $i=1, \ldots, k+l$, for which distance d is available,

$$d(w_p, w_q) + d(w_q, w_r) = d(w_p, w_r). \tag{7}$$

The patterns can be graphically presented as

$$w_1 \underbrace{\cdots}_{d(w_1, w_2)-1} w_2 \underbrace{\cdots}_{d(w_2, w_3)-1} w_3 \cdots w_{k+l-1} \underbrace{\cdots}_{d(w_{k+l-1}, w_{k+l})-1} w_{k+l}, \tag{8}$$

where between the words in the pattern there are empty slots corresponding to the distance between the consecutive arguments in the pattern function, decreased by 1. In Table 1 the distances and patterns for word pairs are presented, where the empty slots are marked by the symbol X, representing the variable in the model, that can be further assigned with different words from the set S. In this example four

patterns are presented for *възраст* (*age*), two patterns for *около* (*about/around/ approximately*) and three patterns for *отговаряща (corresponding).*

Table 1. Example for patterns of word pairs

Distance – $d(w_i, w_j)$	Pattern – $\text{ptrn}(w_i, w_j)$
2	*отговаряща* X *действителната*
2	*отговаряща* X *календарната*
2	*отговаряща* X *паспортната*
1	*около действителната*
1	*около календарната*
1	*възраст отговаряща*
1	*възраст около*
3	*възраст* X X *паспортната*

Applying rule (6), we can generate only a single aggregated pattern (*age corresponding to the passport data*) from those presented in Table 1:

(9) *възраст отговаряща* X *паспортната.*

Some special patterns describe the bigram and trigram "attribute" candidates.

For instance, for *хиперстеничен гръден кош (hypersthenic thorax)* which is filtered from the *N*-grams process due to frequency below thresholds (only 45 occurences) by the distance method we find patterns ptrn(*хиперстеничен, гръден*), ptrn(*гръден, кош*), and ptrn(*хиперстеничен, кош*) with corresponding distances *d*(*хиперстеничен, гръден*) = 1, *d*(*гръден, кош*) = 1, and *d*(*хиперстеничен, кош*)=2, using rule (6), the pattern ptrn(*хиперстеничен, гръден, кош*) can be generated. Actually the inferred pattern cannot be directly added to trigrams set due to the presence of more patterns including *thorax*, like ptrn(*астеничен, гръден, кош*) (*astenic thorax*). The further rules for "attribute-value" extraction show that in such cases *гръден кош* represents the "attribute" with the corresponding values *хиперстеничен* and *астеничен*. The generated patterns and such additional rules for identification of "attribute" bigrams and trigrams from those of "attribute-value" and "values" trigrams and bigrams help us significantly improve the final result.

After collecting the "attribute" candidates we can observe by their frequency values that some PRs do not contain explicit information about them. In such cases it is assumed that "tacit" information means that their condition is in norm and they do not need any special attention. In order to generate more useful patient status structure and to be able to use it for further automatic processing, we also add into the model for such an attribute from the top *N* candidates, the so called "default" values, i.e. "normal".

The extracted "attribute-value" are further checked and analyzed by the experts. To study the correlation of values for different organ characteristics, the medical experts in the project have developed a scale of *normal*, *bad* and *worst*

conditions. Our approach has similarities to the one presented in [2], where the patient smoking status is classified into five categories. Some words from the PRs are chosen as a representative for the corresponding status scale and the other text expressions are automatically classified into these typical status grades. Table 2 illustrates the scales for *limbs* and gives examples for words signalling the respective status. This allows further clustering of the attribute values to these three classes.

Table 2. Limbs characteristics categorisation

Scale	Ankle	Leg	Peripheral Artery Pulsation
0	**normal**	**normal**	normally present
–1	**(light) swelling**	**oedema**	reduced
–2	**solid swelling**	**solid swelling**	**absent**

On Fig. 2 a screenshot of the system containing the dynamically generated structured representation of the patient status is shown. In white colour the assigned values with assigned scale 0 are presented, describing the status in normal conditions The yellow marked values are with a scale –1, i.e., those which are slight variations from the norm. The red coloured values represent data with scale –2, to which special attention should be given because they are indicators of serious complications. Some of the attributes assigned by “default” values are marked in green because they are automatically generated and no explicit information about them is available in the text.

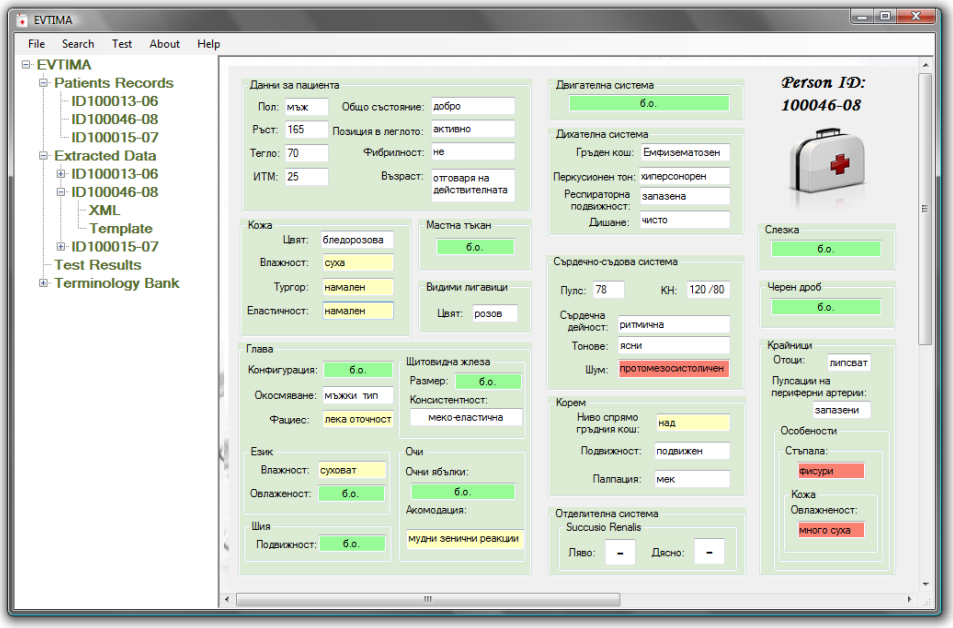


Fig. 2. Screenshot of the system containing dynamically generated structured representation of the patient status

4. Evaluation and results

For our experiments a training corpus is used, containing 1378 PRs and a test set with 6200 PRs. All tests are performed in two modes: processing Patient status sections only and processing full discharge letter texts. Processing the full PRs text is also meaningful, because such "attribute-value" pairs even in another context and with different meaning are available not only in the patient status sections, but also in sections as:

- ***anamnesis*** (personal medical history) – where some current and past patient complaints are described, like hypertension with presented blood pressure values;
- ***laboratory and other tests findings –*** although these are mainly attribute-value pairs, the method cannot be directly applied to this section due to the specific table format representation of the data;
- ***medical examiners comments*** – in this section particular information concerning some specific disorders is presented in more details; for instance, the Ophthalmology examiner can list some information about the glasses dioptres like VOD= 0.6 and VOS=0.6, which can be interpreted as "attribute-value" pairs;
- ***debate*** – this section contains explanations about the deseases development and the patient status changes during the hospitalization.

Tables 3 and 4 present the extracted *N*-grams from the training and test corpus containing 1300 PRs.

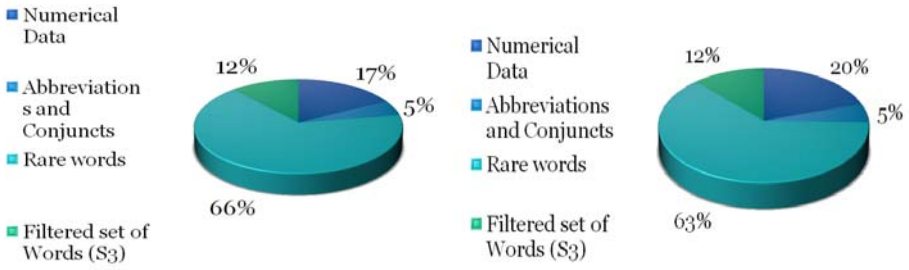
Table 3. Summary statistics for the extracted data for steps 1-6 of *N*-gram method

Set	Values	**Status section only Training set (1378 PRs)**	**Status section only Test set (6200 PRs)**	**Full PR Text Training set (1378 PRs)**	**Full PR Text Test set (6200 PRs)***
Set S – words extracted from PRs	**Total**	169 959	729 893	917 985	3 771 156
	Unique	3159	6178	29 469	69 130
Set Num – Numerical data	**Total**	8857	40 077	149 740	607 807
	Unique	635	1057	8044	23 577
Set CA – Abbrevia-tions and conjuncts	**Total**	35 370	153 796	245 410	1 006 864
	Unique	169	327	1053	1577
Set RW – Words with low frequency	**Total**	5464	11 538	50 828	102 239
	Unique	1972	4078	17 716	38 483
Set S3 – Filtered set S	**Total**	120 268	524 482	472 007	2 054 246
	Unique	383	716	2656	5493

Table 4. Summary statistics for the extracted data for steps 7-11 of *N*-gram method

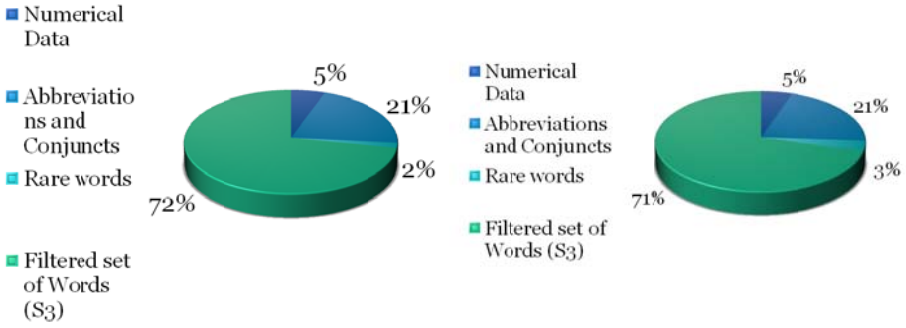
Set	Values	Status section only Training set (1378 PRs)	Status section only Test set (6200 PRs)	Full PR Text Training set (1378 PRs)
Set P – 2-grams	**Total**	73 700	322 167	217 817
	Unique	1542	4720	23 746
Set T – 3-grams	**Total**	43 799	190 998	103 065
	Unique	1830	6180	21 450
Set P1 – Filtered set P	**Total**	22 573	93 025	46 181
	Unique	67	117	279
Set T1 – Filtered set T	**Total**	2146	13 586	2177
	Unique	5	8	7
Set S4 – Filtered set S3	**Total**	59 735	243 809	311 090
	Unique	247	490	2 169

The resulting data (Figs 3 and 4) shows approximately the same distribution of the data in the training and test corpus for the patient status section analyses, both for total word occurrences and for unique words. We can see that the set S3 contains words with high frequency (71% of the corpus) and a small amount of unique words (about 12% of all words in the corpus). This specific structure allows "attribute-value" pairs extraction with high precision using the methods proposed. The "attributes-value" pairs were recognized with 96% precision, using only the *N*-grams method. Filtering rules for trigrams were too strong and only 5 trigram candidates from 2146 in the training set and 8 trigram candidates from 13 586 in the test set meet all the criteria. But not surprising, all of them were correct. The recall was not so impressive (about 87%) due to many attributes presented with lower frequency, word forms and misspelling errors in the PRs text. For unigrams and bigrams the precision is a little bit lower, 92% and 97% correspondingly, due to some complications with common "values" for some "attributes". In contrast with trigrams the recall for unigrams and bigrams is relatively higher – 91%.



(a) 1300 PRs – Training corpus (b) 6200 PRs Test Corpus

Fig. 3. Status sections – unique word



(a) 1300 PRs – Training corpus (b) 6200 PRs Test Corpus

Fig. 4. Status sections – total

Further procedures for collocations patterns and word distance patterns allow us to improve both precision and recalling for "attribute-value" tuples. For the test corpus (1300 PRs) 7183 word pairs patterns were generated, where 720 stable bigrams (with distance 1) and 801 patterns with distance 2 were identified. Using the rules, about 1000 aggregated patterns were generated from them. On Fig. 5 the statistics for the word pair patterns are shown, where the horizontal axis presents the count of generated patterns and the vertical axis presents the distance between words in the pattern. The maximal distance between the words in the generated pattern is 45. Such distances are presented only for singleton words and further those patterns with word distance above 20 are not combined in aggregated patterns.

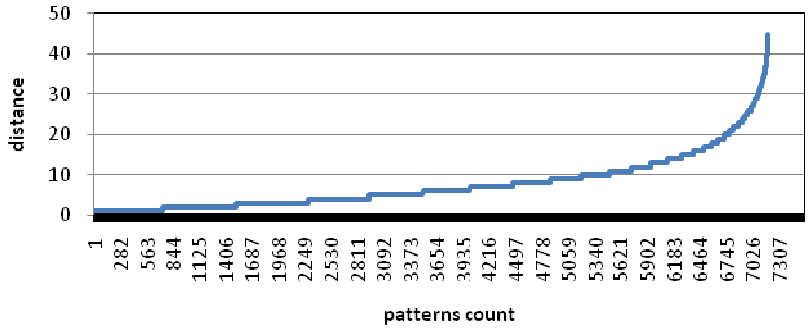


Fig. 5. Summary statistics about the word pair distances and generated patterns in the training corpus

All described methods were implemented in workbench (Fig. 6) that allows us to test different steps results and different methods and rules combinations.

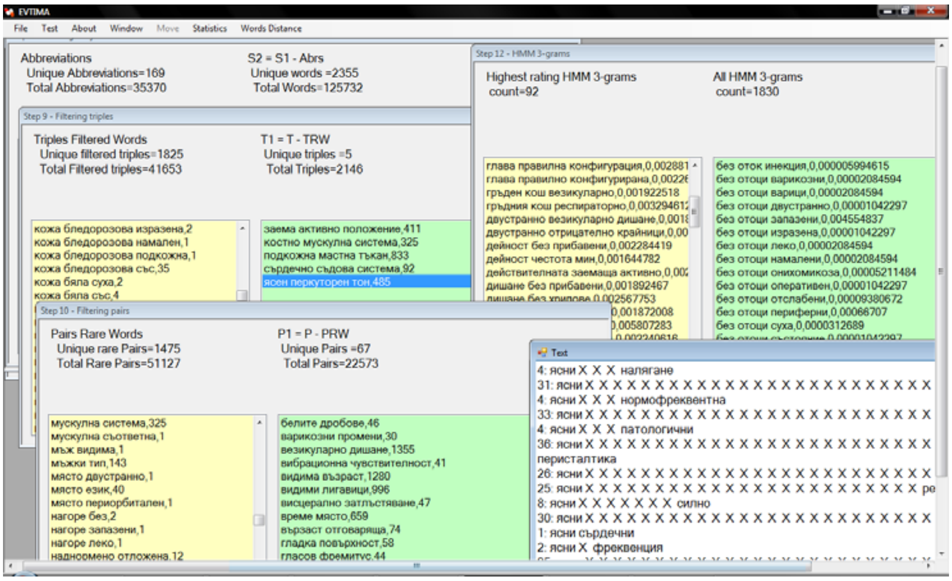


Fig. 6. Screenshot from the workbench for automatic "attribute-value" extraction

The results of automatic extraction of information on the status is relatively high and can be compared with other systems that solve similar tasks, such as system CLEF (CLinical E-Science Framework) that retrieves data for cancer patients [14]; AMBIT that retrieves medical information from biomedical text [15]; MiTAP (MITRE Text and Audio Processing) monitor complications in infectious diseases [16]; caTIES (Cancer Text Information Extraction System) handles medical records [17]; MedLEE (Medical Language Extraction and Encoding System) designed for processing of radiology reports and later expanded to process medical history [18]. Another system is the Mayo Clinic NLP System [13] for structured retrieval patients about their smoking status.

5. Conclusion and further work

This approach shows high precision in "attributes-values" pairs information extraction. The methods discussed are unsupervised and language independent. The approach was also tested for small corpus (about 100) PRs in English language and shows relatively high results – 86% precision. The approach can be also applied for other more complex relations identification in the PRs. Some results for other sections processing were shown.

The presented approach shows that even with lack of resources and difficulties due to mixture of Bulgarian and Latin medical terminology, we can extract certain facts relatively easily, even for attributes with lower frequency in the corpus. The mandatory structure of PRs allows focusing the analyses only in sections containing data of interest. These promising results support the claim that the Information Extraction approach is helpful for obtaining specific medical statements which are described in the PRs texts.

As further work we are planning to add more precise filtering rules. Some methods for abbreviation and word forms processing will be helpful as well. Preprocessing of the corpus for spelling errors correction will the help of some "attributes" frequency increases. Some problems with missing attributes can be resolved by collecting information from PRs with "attribute-value" pairs and predicting from "values" corresponding "attributes".

Further tests in bioinformatics domain of the proposed approach are also planned.

Acknowledgements: The research work is supported by grant No DO 02-292 "Effective search of conceptual information with applications in medical informatics", funded by the Bulgarian National Science Fund in 2009-2012.

References

1. National Framework Contract between National Health Insurance Fund, Bulgarian Medical Association and Bulgarian Dental Association, Official State Gazette No 106/30.12.2005, updates No 68/22.08.2006 and No 101/15.12.2006. Sofia, Bulgaria. **http://dv.parliament.bg/.**
2. Boytcheva, S., I. Nikolova, E. Paskaleva, G. Angelova, D. Tcharaktchiev, N. Dimitrova. Obtaining Status Descriptions via Automatic Analysis of Hospital Patient Records. – In: V. Fomichov, Ed. Special Issue on Semantic IT of Informatica, Int. J. of Computing and Informatics (Slovenia), Vol. **34**, December 2010, No 4, 269-278.
3. Kedar, B., P. Pratim Talukdar, G. Kumaran, O. Pereira, M. Liberman, A. Mccallum, M. Dredze. Lightly Supervised Attribute Extraction. – In: Proc. of the Machine Learning for Web Search Workshop, NIPS, 2007.
4. Poesio, M., A. Almuhareb. Identifying Concept Attributes Using a Classifier. – In Proc. of the ACL Workshop on Deep Lexical Semantics, Ann Arbor, Michigan, June 2005.
5. Marius, P., B. van Durme. What You Seek is What You Get: Extraction of Class Attributes From Query Logs. – In: Proc. of 20th International Joint Conference on Artificial Intelligence (IJCAI'07), Hyderabad, India, 2007.
6. Brin, S. Extracting Patterns and Relations from the World Wide Web. – In: International Workshop on the World Wide Web and Databases, 1998.
7. Agichtein, E., L. Gravano. Snowball: Extracting Relations from Large Plain-Text Collections. – In: 5th ACM International Conference on Digital Libraries, 2000.
8. Wong, Y. W., D. Widdows, T. Lokovic, K. Nigam. Scalable Attribute-Value Extraction from Semi-Structured Text. – In: Proc. of ICDM Workshop on Large-Scale Data Mining: Theory and Applications, 2009.
9. Probst, K., R. Ghani, M. Krema, A. Fano, Y. Liu. Semi-Supervised Learning of Attribute-Value Pairs from Product Descriptions. – In: Proc. of 20th International Joint Conference on Artificial Intelligence (IJCAI'07), Hyderabad, India, 2007.
10. Banko, M., M. J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni. Open Information Extraction from the Web. – In: Proc. of 20th International Joint Conference on Artificial Intelligence (IJCAI'07), Hyderabad, India, January 2007, 2670-2676.
11. Popescu, A.-M., O. Etzioni. Extracting Product Features and Opinions from Reviews. – In: Proc. of EMNLP'2005.
12. Jurafsky, D., J. H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing. Computational Linguistics, and Speech Recognition. Second Edition. Prentice Hall, 2009.
13. Savova, G., P. Ogren, P. Duffy, J. Buntrock, C. Chute. Mayo Clinic NLP System for Patient Smoking Status Identification. – Journal of the American Medical Informatics Association, Vol. **15**, January/February 2008, No 1, 25-28.

14. Harkema, H., A. Setzer, R. Gaizauskas, M. Hepple, R. Power, J. Rogers. Mining and Modelling Temporal Clinical Data. – In: Proc. of 4th UK e-Science All Hands Meeting, Nottingham, UK, 2005.
15. Gaizauskas, R., M. Hepple, N. Davis, Y. Guo, H. Harkema, A. Roberts, I. Roberts. AMBIT: Acquiring Medical and Biological Information from Text. – In: S. J. Cox, Ed., Proc. of 2nd UK e-Science All Hands Meeting, Nottingham, UK, 2003.
16. Damianos, L., J. Ponte, S. Wohlever, F. Reeder, D. Day, G. Wilson, L. Hirschman. MiTAP for Bio-Security: A Case Study. – AI Magazine, Vol. **23**, 2002, No 4, 13-29.
17. Cancer Text Information Extraction System (caTIES). **https://cabig.nci.nih.gov/tools/caties**
18. Friedman, C. Towards a Comprehensive Medical Language Processing System: Methods and Issues. – In: Proc. of AMIA Annual Fall Symposium, 1997, 595-599.