Medical Text Processing for SMDA project

Mikhail Lushnov[1], Timur Safin[2], Maxim Lapaev[3] and Nataly Zhukova[3]

[1] Federal Almazov North-West Medical Research Center, St.Petersburg, Russia,
Professor,
lushnov_ms@almazovcentre.ru,
http://www.almazovcentre.ru
[2] InterSystems Corporation, Moscow, Russia
timur.safin@intersystems.com,
http://www.intersystems.com
[3] Laboratory ISST, ITMO University, St.Petersburg, Russia
m.lapaev@corp.ifmo.ru, nazhukova@mail.ru,
http://www.ifmo.ru

Abstract. Medical care is vital and challenging task as the amount of unstructured and unformalized data has grown dramatically over last decades. The article is dedicated to SMDA project - an attempt to build a framework for semantic medicine application for Almazov medical research center, FANW MRC[4]. In this paper we investigate modern approaches to medical textual data processing and analysis, however mentioned approaches do not give a complete background for solving our task. We spot a process as a combination of existing tools as well as our heuristic algorithms, techniques and tools. The paper proposes a new approach to natural language processing and concept extraction applied to medical certificates, doctors' notes and patients' diaries. The main purpose of the article is to present a way to solve a particular problem of medical concept extraction and knowledge formalization from an unstructured, lacking in syntax and noisy text.

Keywords: Knowledge extraction, Knowledge refinement, Natural language processing, Ontologies, Semantic approach, Medical text processing

1 Introduction

Texts written in natural language have always been an object of tremendous interest for computer science. Appearing to be a rich source of detailed information for humans, hardly can such texts be comprehended and processed by machines as a result of absence of pattern, lack of structure, syntactical noise, term synonymy and ambiguity of terms. A number of approaches to natural language processing emerged since the necessity to deal with such dark data was realized. Most of tools are aimed at solving particular applied tasks, however, still not a single solution meeting all needs and requirements exist. Developed tools and

[4] FANW MRC, URL: http://www.almazovcentre.ru/?lang=en

techniques have spread far outside computer science research applications and are used nowadays in a great number of domain areas. Medicine and medical text analysis is no exception. Our long-term goal is to develop a framework of semantic medical data analysis and a current primary task is to process, analyse and formalize medical certificates and clinical notes for better interoperability on the way to semantic medicine.

2 Current state

Unstructured medical textual data is referred to as an intersection area of natural language and domain-specific terminology and thesauri. Medical text processing is not a new issue, however, it is still a complicated task to extract this data into a well-defined structural storage. The state-of-the-art researches and overviews deal with subjects of automated recognition, machine translation and analysis of clinical records. Still, all presented techniques have a number of restrictions and are far from through processing from raw text to well-structured semantically consistent state. Several attempts of English text processing techniques already exist, but there is an evident lack of methods for other languages, especially for a specific domain area like medicine. Current researches are focused on clinical text segmentation[1] and medical named entity recognition[2], [3] solutions which are of high value and appeared to be efficient but for the one drawback: the techniques are applicable only for languages they are targeted at.

Another direction of investigations in medical text processing, which is of great importance for us, is medical notes refinement from the point of view of text correctness and consistency. Manual analyses has shown that clinical records are mistype-prone pieces of textual data with a perceptible number of hard-to-identify-and-interpret-automatically contractions and acronyms, which assumes normalizing and bringing text to a conventional purified plain form. Researches suggest context-aware spell-check[4] and expanding acronyms and abbreviations[5]. Nevertheless, proposed techniques abut the restriction: correction of spelling errors is a language-specific task, whereas de-acronymizing demands a voluminous dictionary.

Other recent subjects under research and projects are concerned with pharmacy and focus on drug interactions[6] and drug reactions[7] as well as machine translation of medical texts and standardization[8],[9] which appear to be a challenge as a result of peculiarity and specificity.

As far as Almazov medical research center is concerned, all textual data is processed manually as a result of unavailability of processing and analytical tools for text written in Russian (a language having complicated spelling rules, almost free word order within the sentence, a dominance of prefix-suffix-formed lexemes, rich synonymy). Existing and used in the establishment information system does not provide many options to deal with text.

3 Input data description and problem statement

All medical data of Almazov medical center is presented by objective numerical data (analysis and examination indicators provided with measurement units), organoleptic test indicators (colour, smell, transparency and others) and textual subjective notes which are of the highest interest for us in this study. A great amount of data, both historical and current, should also be taken into consideration. Statistical information on overall volume is presented in Tab. 1 and reflects more textual data rather than other data-types. Notes include life or disease anamneses, ambulatory, assignment or permanent diagnoses, patients' diaries and complaints, recommendations, prescriptions, surgery protocols and tens of other types.

Table 1. Overview of textual data in medical research center

Amount of data for one patient (from .. to)	**1 .. 300 notes**
Average amount of data per patient	**100 notes**
Average number of patients per doctor per day	**20 notes**
Amount of doctors in the center	**400 doctors**

Manual thorough textual data analysis has shown that medical texts are specific domain-area related text pieces possessing a number of traits which leads to lack of possibility to treat them as usual textual data by widely used categorization, classification and scraping tools: 1) completely verb-absent statements or text with a lack of verbs; 2) a great number of contractions and designations; 3) frequent misspellings, lack of syntactical structure. However, not only do medical text pieces have a number of disadvantages from automated processing point of view, they also have a predictable (and, thus, templatable) style and structure: 1) closed and well-defined domain area with its dictionaries and references and established writing style; 2) declarative medical certificates' and notes' style and language; 3) established lexemes and collocations.

Clinical notes dramatically differ from a normal text as a result of particular job features such as restricted time for one patient, focusing on diagnosis, but not the written text and settled word reductions within the domain area which leads to hard-to-recognize-and-process noise and other inconsistencies. To sum up and state the problem, no unique, single and efficient at the same time tool considering peculiarity and specificity of domain area yet exists, thus, an appropriate tool-chain as a composition of existing tools and developed specific domain-sensitive tools and techniques is required. The tools must be aimed to tasks of noisy text pre-processing, concept extraction and syntactic and semantic analysis based on extracted establishment-specific dictionary.

4 Medical textual data processing

Medical certificates and doctors' notes are specific pieces of textual data rich in terminology and descriptive lexemes which require careful treatment and multi-stage processing techniques including heuristic algorithms, concept extraction and matching and syntactical parsing. Text analysis includes stages of syntactical analysis, tokenization and conceptual analysis afterwards which requires syntactical validness to be minded (i.e. full stop sign at the end of the sentence, word order and other conditions). Our general strategy includes the following steps: 1) organization-sensitive context heuristic and regular expression-based text analysis and clean-up to convert text pieces into a conventional form and to fit the requirements for input of next preprocessing and processing tools; 2) concept and entity analysis, extraction, indexing and matching to build a dictionary within the context of organization (a medical establishment in described case); 3) syntactical parsing, analysis and rule- and template-driven formalization. In order to verify applicability of analyser for particular tasks in medical text processing we provided the syntactical parsing tool with a variety of pieces of text related to domain area and checked the results which turned out to be consistent.

4.1 Heuristic and regular expression-based analysis

Text analysis and overview has shown that the following inconsistencies may occur in the text (a bright example is "alcohol Cons-n is rare and in min. quant." which is to be converted into "Alcohol consumption is rare and in minimum quantities."):

1. not all meaningfully complete sentences end with "full stop" token;
2. "full stop" token may be encountered in the middle of the sentences in cases of word contractions or as a mistype instead of comma;
3. mistypes as a result of time restrictions may occur within the text;
4. some words and widely-spread within the domain area collocations are contracted sometimes;
5. line break instead of "full stop" sign in some sentences;
6. line break in the middle of the sentence;
7. words and sentences in capitals (activated CAPS LOCK key on the keyboard);
8. absence of spaces between lexemes and after punctuation tokens;
9. orthographical mistypes within the word;
10. extra spaces, absence of spaces and absence of "full stop" tokens at the end of the sentences.

Syntactical structure and punctuation are vital for correct interpretation of phrases and the following structural analysis. All of the mentioned above text defects necessitate the following preprocessing filters:

1. trimming extra spaces and space insertion if missed;

2. heuristic use of orthography check and correction services' dictionaries to get rid of mistypes;
3. removal of extra spaces, line breaks and other noise by means of regular expressions;
4. heuristic analysis and word complementing by means of free third-party services and technologies as well as removal of "full stop" sign as a signal for contracted words;
5. first letter capitalization in cases of a word followed by a "full stop" token by means of templates based on regular expressions;
6. heuristic conversion of letters to lower case for all symbols of text except the ones at the beginning of the word;
7. complementing the text by "full stop" token in case the next word starts with capital letter with a help of regular expressions.

As soon as all mentioned above filters are applied we have an almost cleaned-up text with minimum of noise which may be used as a source text for concept and entity extraction.

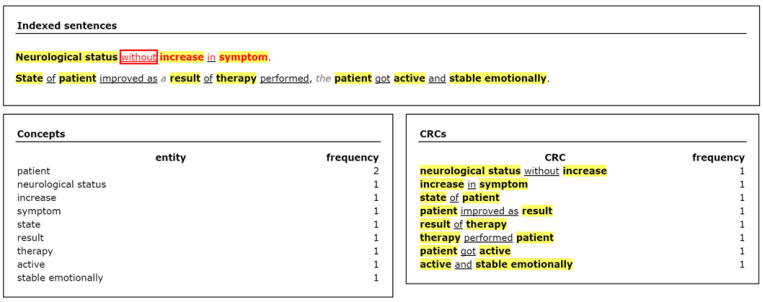
4.2 Concept and relation extraction with iKnow

InterSystems iKnow[10],[11],[12] is the technology used in Caché database for extraction of structured information from unstructured sources. iKnow uses set of common heuristics which simplify the process of natural language processing in the source languages and makes this process as fast as possible, providing good enough precision. The tool is language-agnostic, thus, it could handle wide set of input natural languages, including English, German, French, Dutch (Flemish), Russian, Ukrainian, Swedish or Japanese. Moreover, iKnow supports external stemming engines for languages with complex morphologies (i.e. Russian, Ukrainian). iKnow engine splits input text into sentences, which are divided by punctuation tokens ("full stop" in example). Afterwards it analyses sentence using language-specific model heuristics. The result of analysis is a sequence of Concepts and Relations, built set of CRC triplets (concept-relation-concept), and a set of various indices based on statistical metrics. Relation here is a word or a group of words connecting two concepts by a specific relation. Lets try explain basic concepts of iKnow in a visual form, please consider this simple input text: "Neurological status without increase in symptoms. State of patient improved as a result of therapy performed, the patient got active and stable emotionally."[5]. After indexing in iKnow we get the result (Fig. 1).

iKnow semantics analysis recognizes key elements such as concepts (yellow-coloured), relations (underscored), non-relevant words (italics), negations (red-coloured).

Taking into account that relations are commonly verbs, and nouns with adjusting words are concepts, and the fact that number of verbs are relatively

[5] Note that we present examples in English for comprehensibility reasons with no loss of meaning despite the fact that source text is written in Russian



Indexed sentences

Neurological status without **increase** in **symptom**.

State of **patient** improved as *a* **result** of **therapy** performed, *the* **patient** got **active** and **stable emotionally**.

Concepts

entity	frequency
patient	2
neurological status	1
increase	1
symptom	1
state	1
result	1
therapy	1
active	1
stable emotionally	1

CRCs

CRC	frequency
neurological status without **increase**	1
increase in **symptom**	1
state of **patient**	1
patient improved as **result**	1
result of **therapy**	1
therapy performed **patient**	1
patient got **active**	1
active and **stable emotionally**	1

Fig. 1. Results of concept extraction by means of iKnow

stable in any given language, simple approach used at the core of iKnow engine provides very good performance: it starts analysis from relations boundaries, then it extends horizontally using language model and its morphology heuristics. Moreover, such generic approach provides fast start in iKnow enabled application development you could get reasonable indexing results even without application of any domain-specific ontology.

Worth to mention that cleaner text is important for such language-oriented analysis (it helps in any approach of semantic analysis though, but is essentially important for the case of generic, domain agnostic language model). That is why, before processing of text in iKnow it is important to clean input text from local abbreviations and any possible short-cuts (see Heuristic and regular expression-based analysis above). iKnow itself could be used for cleaning up the input data (using text converters, like html converter, or dictionary matching for abbreviation expansion) but this not reduce importance of cleanup heuristics as described above. iKnow uses domain separation for the separate data source indexed with their own set of settings (source, language, stemming, etc.). Once rebuilt iKnow domain information could be used in various scenarios for analysis:

- adhoc lookups using iKnow query API for retrieval of any kind of relationships between concepts, CRC or CC matches;
- Text Categorization framework for automatic labeling of natural language text with predefined set of categories, using naive Bayes, vector machines, decision trees and other related algorithms;
- data extracted using iKnow may be natively projected to the InterSystems DeepSee for visualization inside of business analytics platform, or easily embedded into modern web applications using their REST projection.

Efficiency of iKnow engine implementation and its native integration with the multi-model database platform Caché allows applying wide set of semantics analysis and processing techniques right inside database management system (DBMS) environment without any extra overhead. Based on iKnow processing we acquire a set of medical concepts which are entries to form a domain area

organization-sensitive dictionary, which are later processed with a help of domain experts to be included in further processing stages.

4.3 Syntactical analysis

After preprocessing and conceptualization stages a syntactical analysis stage is desired for further natural language processing and formalization which requires a powerful rule-based tool with support of text source language and manageable output formats (as a result of iKnow using verbs to split the sentence in Russian into concepts, and we mentioned that clinical records lack in verbs). One of the tools to fit the task is SemSin - a semantic and syntactical analyser of texts written in Russian [13],[14], the tool widely used in our projects for particular linguistics analysis tasks.

The tool fulfils morphological and syntactical analysis of sentences written in Russian and builds the dependence tree (Fig.2) for it. For the description of morphological, syntactic and semantic characteristics of words the dictionary produced by InterSystems iKnow at previous stage and the classifier of lexemes are used. Analyzer functioning is carried out by means of a set of production rules. Feature of rules is the decision making about an establishment of syntactic links to simultaneous removal of a morphological ambiguity. SemSim takes a piece of text (a sentence or a paragraph) as an input data and tokenizes it. As soon as tokenization is completed, the text undergoes a morphological analysis stage. Each token is supplied with morphological, syntactical and semantic information. The next step includes processing by a set of production rules for general language and a specific set of rules suggested by experts in medicine and adapted for medical texts to disambiguate the word chain and translate linear syntactical structure into a syntactical tree. SemSin attempts to check whether the token matches a rule and to carry out processing acts if it does.

A significant trait of the tool is ability to work only with grammatically and orthographically correct sentences, which, however, does not pose any problems for us as input text is already spell-checked and refined. The outputs of the analyser are sentence tree visualized as a graph. Another and the primary product of the tool is a structurally valid XML describing presented graph which may later be parsed to scrape data relevant for the task and make it possible to organise a more precise text processing, concept extraction and ontological matching pipeline.

4.4 Pattern analysis

After doctor's note is preprocessed, cleaned-up and parsed as a sentence tree, further processing and analysis are possible based on thesaurus or references and analysis patterns. Well-established syntactical structure and widely used cliches in medical certificates and doctors' notes make it possible.

A well-documented and visual Drools engine as tool to draw up rules was used so that experts in domain area may easily create and edit patterns. Rules process an input XML produced by SemSin to search for keyword entries and execute

specific rules if pre-defined word sequences are encountered. Specific rules follow relational tree from a keyword as an entry point to define what is related to the keyword and what kind of relation it has, for example, "is-a", "has-a", "how", "why" and so forth. Step-by-step deepening gives a full image of patients' state. For instance, we analyse example from Section 4.2. Its structural tree as a result of SemSin processing is presented in Fig.2.

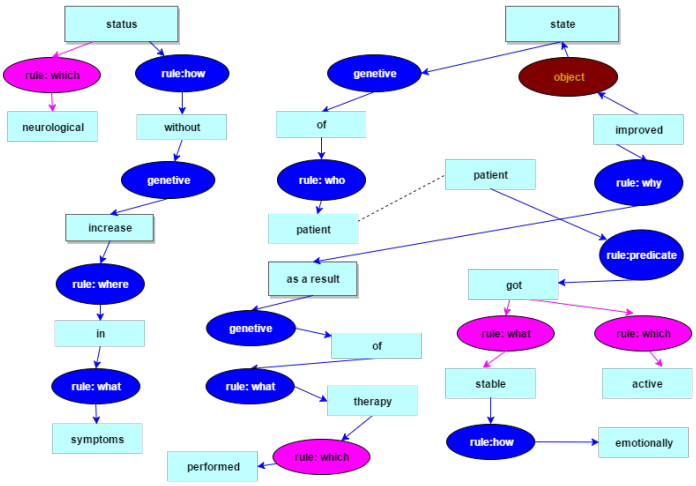


Fig. 2. Structural-syntactical tree of example medical text pieces

In provided example, the word "status" following the cliches is addressed to patient's status at the moment of visit. Thus, searching for this lexeme we may investigate node relations to clarify it: following the branch "which" we may conclude, what kind of status ("neurological" in our case) patient has; the same way, following the branch "how" and successive branches we come to a conclusion that there is no increase in symptoms which strictly describes state between possible variations "with increase", "without increase". A rule pseudo-code (Alg. 1) to find out the reason of state improvement is presented for the second example sentence. Provided that language grammar has a finite set of structure-dependent rules and morphemes, especially ones related to a specific domain area, which, in turn, has a finite number of cliches, sentence analysis is amenable to formalization involving expert knowledge.

Algorithm 1 A rule pseudo-code for example sentence

```
rule F1
when keyword = "state" then predicate = traverseBranch("object")
  when predicate.hasRelation("why") then processRelation()
  end end
```

5 SMDA architecture overview

Described above text processing approaches are integrated into a powerful semantic medical data analyses and management system - SMDA. The system is a framework represented by a number of layers divided according to the principles of data structuring level and processing techniques: unstructured multimodal raw data; preprocessed semi-structured data; formalized and structured data and knowledge; knowledge graph; and, finally linked data space. A variety of techniques and heterogeneousity of data at every stage necessitate a highly distributed architecture with a number of stores for each type of data such as an object data base for raw and preprocessed data as well as triple store for structured and refined knowledge with a REST or SOAP interaction interfaces. Having a highly distributed architecture the system requires endpoint services' unification to add scalability and reduce time consumption of development, design and debug stages in conditions of dealing with big data. In cases of protocol or format incompatibilities adapters are to be embedded in the pipeline. The framework assumes a tool-chains and workbenches to be integrated for particular processes and storage tasks (Fig. 3) and a process pipeline and component interaction for text processing in particular (Fig. 4).

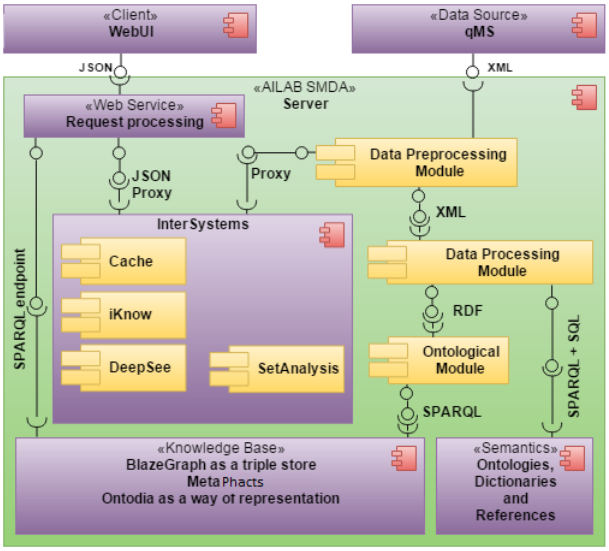


Fig. 3. SMDA component interaction

– mentioned above text processing techniques including both our technologies and third-party services for text clean-up to fit the format;
– InterSystems iKnow as a tool to deal with unstructured data by dividing text into relational and associated entities as well as indexing tasks;

– SemSin tool for syntactical and semantic analysis and relation extraction tasks;
– a number of tools to deal with raw numerical data or numbers extracted from pieces of textual data (OLAP-technologies to optimize access, statistical and analysis tools and AILAB set of processing algorithms);
– a set of repositories to store and access both formalized and unstructured data supporting a sufficient volume of data for the task (over 5 billion triples for triple-store) (BlazeGraph[6] and InterSystems Caché, respectively);
– inference and representation mechanisms (MetaPhacts[7], AILAB Ontodia[8], rule-driven module designed by the laboratory and SPARQL-endpoints).

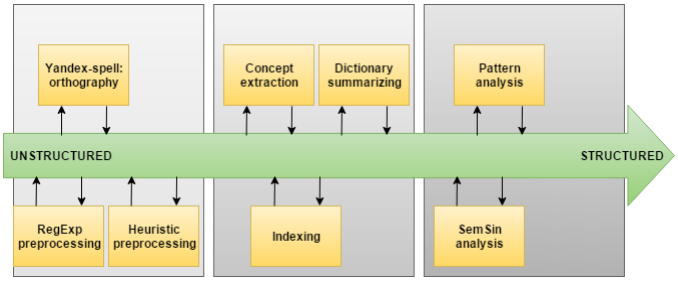


Fig. 4. Textual data processing pipeline

Presented text processing pipeline is aimed at taking plain medical textual notes as an input and producing a set of triples as an output.

6 Validation and verification

Presented pattern approach needs to be evaluated, thus, we developed a tool based on SEO-analysis services and produced a number of templates based on frequent sequences all over the text of sample clinical notes and verified the results.

6.1 Example and expected results

In order to inspect the approach thoroughly and in a more accurate way, we excluded common patterns such as prescription-like notes, enumerations of "indicator" - "value" pairs and other similar obvious templates. A more branching set of patterns was analyzed. We expect all phrases matching the pattern to be found and processed. The most frequent templates within the sample text are the following patterns (a few of them, Fig. 5):

[6] BlazeGraph, URL: https://www.blazegraph.com
[7] MetaPhacts, URL: http://aksw.org/Partner/Metaphacts.html
[8] Ontodia, URL: http://www.ontodia.org

1. if any word in a set "disease", "disorder", "affection" is encountered, it has two structural branches having "of what" and "which" relations and provides information on which organ and in what degree is affected;
2. whenever either "system" or "organs" words occurs, structural branches refer to "what" or "which" relation pointer to specific organism part followed by a "column" token meaning "is_a" relation with enumeration of "parameter" - "state" pairs afterwards;
3. when a word sequence "according to the results" is found, the phrase, containing the sequence, branches into "of_what" relation(test, examination) and "what" relation (enumeration of parameter-state pairs).

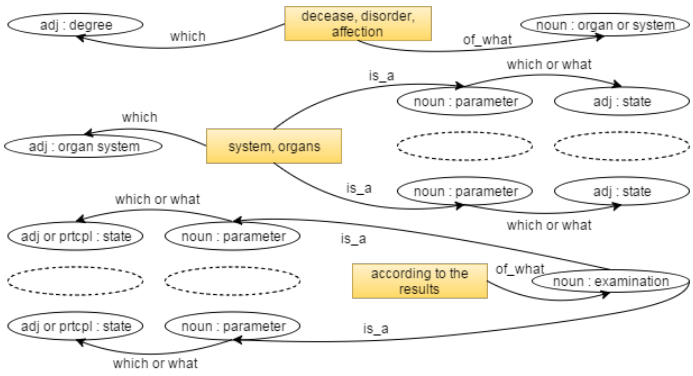


Fig. 5. Structural patterns based of frequent sequences

6.2 Evaluation

To verify reliability of the techniques a set of rules representing selected patterns on sequence frequency basis (often encountered patterns) was applied to 3085 sample clinical notes to check whether rules produce exactly what was intended to extract. Results of processing (all phrases matching the patterns) were analyzed carefully, domain area experts were engaged to conduct a more authentic evaluation. All matched sentences turned out to have been parsed with a high degree of confidence which indicates conditional applicability of proposed techniques. We present a couple of phrases for some patterns used during the evaluation process as examples:

1. "Varicose affection of lower extremity veins", "Atrial fibrillation, a transient form of the constant disorder of atrioventricular conduction";
2. "Digestive organs : abdomen at superficial palpation is soft and smooth, liver not enlarged, liver edge is moderately dense", "Respiratory system: vesicular breathing";

3. "According to results of MRI: consequences of lacunar CVA of left cerebral artery of unknown period", "According to results of dopplerography of brachiocephalic artery: hemodynamically significant stenoses are not revealed".

7 Conclusion and future work

In this study we showed that introduced techniques are applicable for the issue of medical clinical notes analysis and semantic structuring. Along with the techniques we introduced a pre-processing tool-chain that dramatically improves the quality of texts from point of view of consistency and conventionality which makes further analysis and parsing processes at next stages of the tool-chain possible. Validation and verification has shown that using the introduced techniques for parsing and concept extraction rather than manual dictionary formation sufficiently decreases time consumption required for entity acquisition dealing with large volumes of data and may be inlined into a processing pipeline of Russian medical text analysis for semantic medicine applications. Despite a fare sufficiency of introduced techniques further improvements are desired, which include more precise clean-up algorithms, more flexible pattern combinations and a greater amount of patterns to extract as many semantically-structured data as possible, which are the matter of further study, analysis and work.

References

1. G. Orosz, A. Novak, G. Proszeky, "Hybrid text segmentation for Hungarian clinical records", Advances in Artificial Intelligence and Its Applications, 2013, vol. 8265, pp. 306-317
2. M. Chernyshevich, V. Stankevitch, "IHS-RD-BELARUS: clinical named entities identification in French medical texts", Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum (2015).
3. Alanazi, Saad, Bernadette Sharp, and Clare Stanier. "An evaluation of AMIRA for named entity recognition in Arabic texts." (2015).
4. S. Borbala, A. Novak, G. Proszeky. "Context-aware correction of spelling errors in Hungarian medical documents", Computer Speech & Language, 2016, vol.35 pp.219-233.
5. S. Moon, S. Pakhomov, N.Liu, J. O Ryan, G. B Melton, "A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources", J Am Med Inform Assoc, 2014, vol.21(2) pp. 299-307
6. M. Rastegar-Mojarad, R. D. Boyce, R. Prasad, "UWM-TRIADS: classifying drug-drug interactions with two-stage SVM and post-processing", Proceedings of the Seventh International Workshop on Semantic Evaluation, 2013, vol.2, pp. 667-674
7. F. Stefanie, H. Dalianis, "Adverse drug event classification of health records using dictionary-based pre-processing and machine learning." 6th international workshop on health text mining and information analysis (LOUHI), 2015
8. W. Krzysztof, K. Marasek, "Polish-English statistical machine translation of medical texts", New Research in Multimedia and Internet Systems. Springer International Publishing, 2015, vol.314, pp. 169-179.

9. L. Deleger et al., "Building gold standard corpora for medical natural language processing tasks." AMIA. 2012.
10. Brands, Michael Rik Frans, and Dirk Medard Helena Van Hyfte. "Data integration and knowledge management solution." U.S. Patent No. 7,428,517. 23 Sep. 2008.
11. Brands, Michael Rik Frans, and Dirk Medard Helena Van Hyfte. "Data processing based on data linking elements." U.S. Patent No. 7,912,841. 22 Mar. 2011.
12. Brands, Michael Rik Frans, and Dirk Medard Helena Van Hyfte. "Data analysis based on data linking elements." U.S. Patent No. 9,053,145. 9 Jun. 2015.
13. Boyarsky, K., Kanevsky, E.: "The semantic-and-syntactic parser SEMSIN". In: International Conference on Computational Linguistics Dialog-2012 (2012).
14. D. Mouromtsev, L. Kovriguina , Y. Emelyanov, D. Pavlov, A. Shipilo, "From spoken language to Ontology-Driven Dialogue Management", Lecture Notes in Computer Science, 2015, vol. 9302, pp. 542-550