ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ ИЗ КЛИНИЧЕСКИХ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

Шелманов А. О. (shelmanov@isa.ru),
Смирнов И. В. (ivs@isa.ru)
Институт системного анализа Российской академии наук, Москва, Россия

Вишнёва Е. А. (vishneva@nczd.ru)
Научный центр здоровья детей, Москва, Россия

Ключевые слова: обработка клинических текстов, извлечение информации, клинические тексты, извлечение атрибутов заболеваний, медицинские тексты

INFORMATION EXTRACTION FROM CLINICAL TEXTS IN RUSSIAN

Shelmanov A. O. (shelmanov@isa.ru),
Smirnov I. V. (ivs@isa.ru)
Institute for Systems Analysis of Russian Academy of Sciences, Moscow, Russia

Vishneva E. A. (vishneva@nczd.ru)
Scientific Centre of Children Health, Moscow, Russia

We present and evaluate the pipeline for processing of clinical notes in Russian. The paper addresses the tasks of drug identification and disease template filling, which are related to entity recognition and relation extraction. The disease template filling consists in recognition of disease mentions in text, mapping them to concepts of a thesaurus, and discovering their attributes. Discovering attributes means identifying corresponding spans in text, linking them to diseases, and normalizing them i. e. determining their generalized meaning from a predefined set. We implemented tools for determining the following attributes of disease mentions: negation; the flag indicating the disease mention is not related to a patient; severity; course; and body site. For different tasks, we used different techniques: rule-based patterns and several supervised machine-learning methods. Since there were no annotated corpora of clinical notes in the Russian language available for research purposes, we annotated a dataset, which we used for training and evaluation of the developed tools. The created corpus is available for researchers through the data use agreement.

Keywords: clinical text processing, information extraction, annotated corpus, clinical narrative, disease template filling, medical text, EHR

1. Introduction

A vast amount of clinical data is stored as free text. Electronic health records of medical facilities accumulate radiology, echocardiography, and electrocardiogram reports, anamnesis, results of ultrasound diagnostics, discharge summaries, and many other types of notes related to patient healthcare history that are written in a natural language. This is a very rich knowledge source, which is still difficult to exploit because of its unstructured nature. Reworking it into computable form can benefit the biomedical research, patient medical history management, and eventually improve the healthcare. Although there are many natural language processing techniques developed for information and knowledge extraction, the specificity of clinical narrative and tasks arising in the medical domain facilitate the development of specialized methods, language resources, and tools. It is a promising and fruitful scientific direction in natural language processing.

Much of the research in this direction is focused on processing English clinical texts. However, it is also important to create tools and resources for other languages. In the current research, we present and evaluate the pipeline for processing of clinical notes in Russian. The paper addresses the tasks of drug identification and disease template filling, which are related to entity recognition and relation extraction. The disease template filling consists in recognition of disease mentions in text, mapping them to concepts of a thesaurus, and discovering their attributes. Discovering attributes means identifying corresponding spans in text, linking them to diseases, and normalizing them i. e. determining their generalized meaning from a predefined set. We implemented tools for determining the following attributes of diseases: negation; the flag indicating the disease mention is not related to a patient; severity; course; and body site. For different tasks, we used different techniques: rule-based patterns and several supervised machine-learning methods. Since there were no annotated corpora of clinical notes in the Russian language available for research purposes, we annotated a dataset, which we used for the training and evaluation of the developed tools. The created corpus is available for researchers through the data use agreement.

The rest of the paper is structured as follows. Related work is reviewed in section 2. In section 3, we discuss the developed methods and tools. Section 4 describes the annotated corpus. In section 5, the experiment results are described. Section 6 concludes and outlines the future work.

2. Related work

Natural language processing of medical texts is a rapidly developing research area. A big number of challenges conducted in the last few years that are devoted to the problems of clinical and biomedical information retrieval and text processing reflects the growing interest in this area for academic community. We briefly review some of them related to information extraction from English clinical texts. 2012 i2b2 /VA Challenge was focused on temporal relation extraction from clinical narratives (Sun et al., 2013). Pilot ShARe/CLEF eHealth 2013 Evaluation Lab set tasks of identifying in clinical

reports disease mentions and acronym abbreviations as well as mapping them to thesaurus (Suominen et al., 2013). On ShARe/CLEF eHealth 2014 Evaluation Lab these tasks were extended to disease template filling—identification of disease mentions and their attributes (severity, course, body site etc.). SemEval 2014 Task 7 was similar to CLEF eHealth 2013 and was aimed at disease and acronym/abbreviation identification and mapping to thesaurus[1]. Sem Eval 2015 Task 6 was about temporal information extraction from clinical texts[2]. SemEval-2015 Task 14 was similar to CLEF eHealth 2014 disease template filling task[3]. We should also note CLEF eHealth 2015 challenge that although is not directly related to clinical text processing, but shows that efforts are not limited to medical information extraction for English language. CLEF eHealth 2015 shared task introduces the challenge of named entity recognition in French biomedical texts[4]. This initiative became possible because of recent creation of French annotated biomedical corpus Quaero (Névéol et al., 2014).

The problems of processing clinical notes in Russian we focus on in the current work are largely similar to the tasks of entity recognition and disease template filling introduced by the CLEF eHealth 2014 challenge. The majority of the CLEF eHealth participants applied to this task both the rule-based approaches and supervised machine-learning techniques for extracting different attributes, e. g. (Hamon et al., 2014; Liu and Ku, 2014; Huynh and Ho, 2014). (Ramanan and Nathan, 2014) applied to this task purely rule-based methods. In (Mkrtchyan and Sonntag, 2014), approach based on deep dependency parsing was implemented. Several participants adapted and expanded the cTAKES (Savova et al., 2010) medical text-processing framework, e. g., (Sequeira et al., 2014; Johri et al., 2014). The participants commonly exploited the MetaMap (Aronson and Lang, 2010) and NegEx (Chapman et al., 2013) tools for mapping found terms to thesaurus concepts and negation detection correspondingly.

Besides, the problem of disease template filling was recently addressed in (Dligach et al., 2014). The severity and body site attributes were linked to disease annotations with SVM classifier. The researchers tested several kernels including the novel tree-based kernel, and found that the common radial-based kernel was suitable for the aforementioned tasks. The evaluation against rule-based baselines showed the advantage of machine learning techniques.

Since there was no basic toolchain for processing clinical texts in Russian, we had to create it from the scratch. Our medical term (diseases, drugs, symptoms, body sites) identification and normalization module is an analogue of MetaMap. The developed module for disease negation detection and the module for detection that disease is not related to a patient apply approaches implemented in NegEx. Although our modules are based on somewhat similar approaches to those implemented in the tools for English text processing, we made some modifications in them that take into

---

1 http://alt.qcri.org/semeval2014/task7/

2 http://alt.qcri.org/semeval2015/task6/

3 http://alt.qcri.org/semeval2015/task14/

4 https://sites.google.com/site/clefehealth2015/

account peculiarities of the Russian language. In modules for discovering severity and course attributes of disease mentions as well as for linking body sites to disease mentions, we implemented the state-of-the-art machine learning techniques.

Supervised machine learning methods are widely used in natural language processing. To apply them to clinical text processing it is necessary to create annotated corpora for the clinical domain. We are currently developing such corpus for Russian language. It contains clinical free-text notes annotated with treatment, symptom, drug, body sites, disorder / disease mentions, their attributes and relations. The closest analogous for English are the corpus of the Shared Annotated Resources (ShARe) initiative (Pradhan et al., 2015) and the corpus of Strategic Health IT Advanced Research Project: Area 4 (SHARPn)[5].

3. Methods for processing of clinical texts in Russian

The developed clinical text-processing pipeline begins with basic NLP analysis. For tokenization, part-of-speech tagging, sentence splitting, and lemmatization we used well-known pipeline for Russian from AOT.ru (Sokirko, 2001). The dependency syntactic parsing was performed by MaltParser (Nivre et al., 2007) trained on SynTagRus (Apresjan et al., 2005) with configuration described in (Nivre et al., 2008; Sharoff and Nivre, 2011; Smirnov et al., 2014). We find the performance of both of these tools on clinical texts satisfactory. The next step in the pipeline is medical term identification and normalization. It is followed by the disease/symptom negation detection and determining whether disease mentions are related to a patient or not. We will refer to the latter task as "not patient" flag detection. The pipeline concludes by discovering severity and course attributes of disease mentions and linking body sites entities to disease mentions.

3.1. Identifying medical terms in text and mapping them to thesauri

To solve many tasks of clinical and biomedical text processing it is necessary to perform the basic identification of medical terms in text, term normalization, and mapping them to semantic types. For these tasks, we developed a heuristic- and thesaurus-based module. We used the module in the conjunction with the two thesauri: UMLS Metathesaurus (Schuyler et al., 1993) for disease, symptom, and body site identification; a thesaurus based on State Register of Drugs (SRD)[6] for the drug identification.

UMLS Metathesaurus is an extensive compendium of medical lexicons, classifications, code sets, and thesauri. The main feature of it is that the concepts with the similar sense from different knowledge sources are mapped to the single CUI (concept unique identifier) in the metathesaurus. Thus, it is possible to match concepts

[5] http://informatics.mayo.edu/sharp/index.php/Main_Page

[6] http://grls.rosminzdrav.ru/

from any of the sources with each other and unify them in a single system. Among many other knowledge sources, UMLS Metathesaurus incorporates MeSH (Lipscomb, 2000), SNOMED-CT (Bos and Donnelly, 2006), ICD-10[7]. UMLS also provides Semantic Network that maps concepts into one or more coarse-grained semantic types (McCray, 1989). UMLS makes a big contribution to many kinds of medical information-analytical systems and especially to medical NLP systems because it helps to find in text different representations of terms, normalize them to CUIs, and determine their semantic types. Although it contains plenty knowledge sources, the only Russian thesaurus present in UMLS is MeSHRUS[8]. ICD-10 was also translated to Russian[9], but Russian version was not directly integrated into UMLS. MeSHRUS contains about 27,000 concepts and 85,000 terms. It covers many semantic types, but it does not contain drugs and lacks many medical procedures.

SRD is a database of all drugs officially registered and allowed for sale in Russia. Among other information, drugs in SRD have a trade name, short definition, and an international name of an active chemical. Since many drugs have different trade names but almost the same compound and can substitute each other in medication, we preprocessed the database and grouped drugs with similar active chemicals into concepts of thesaurus. We got 3,600 unique concepts and almost 12,000 terms. The result thesaurus can be used to find in clinical texts mentions of similar drugs even though they have different names.

One of the main problems of using thesauri for information extraction is that free text contains usually much more term spellings than the thesauri. It makes simple strict matching often ineffective in terms of recall and forces to invent fuzzy approaches. The problem is especially crucial for medical domain because of the high variability of medical term spellings.

The method implemented in our module was mainly inspired by MetaMap (Aronson and Lang, 2010)—a linguistically motivated tool for mapping terms from medical texts to concepts in UMLS Metathesaurus. Many modern text-processing systems in the medical domain use MetaMap in their pipelines for term identification and normalization (e. g., cTAKES (Savova et al., 2010)). However, this tool can be applied only to English text because it strictly relies on handcrafted heuristics and English lexis (Aronson and Lang, 2010). The developed module for processing texts in Russian is also rule-based and adopts the similar approach implemented in MetaMap. It generates extensive amount of term variants from text expressions, performs fuzzy comparison between the variants and the thesaurus terms, then ranks the variants by heuristically reasoned score, and picks the most confident ones.

In the first step, the parser scans the text for single keywords from the thesaurus using its reverted index, which is preliminarily built by mapping lemmatized tokens of thesaurus terms to thesaurus concepts. Stop-words like particles, punctuation, prepositions are pruned from the index. The found keywords become the seed term variants for the second step—generation of complex variants.

---

[7] http://apps.who.int/classifications/icd10/browse/2015/en

[8] http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MSHRUS/

[9] http://apps.who.int/iris/handle/10665/87721

The generation procedure constructs new term variants from a keyword by expanding it with nodes of the syntactic tree that are below or above the keyword and with words from linear context around the keyword. The procedure is driven by heuristics and constrained by parser parameters (maximum syntax tree depth, maximum number of nodes above the keyword, window of the linear context etc.). Heuristics check whether a new variant should be generated on each step of the expansion. They prune variants with dangling conjunctions and punctuation, splitted complex prepositions, and other minor cases.

The third step is the assessment of how the generated variants correspond to the thesaurus terms. For each token in the variant, the parser determines the sets of thesaurus terms that correspond to the token and unite them in a single set. Then parser assesses the similarity between the variant and terms from the set. The assessment is a value between 0 (the weakest match) and 1 (the strongest match), which is computed as a linear convolution of three components: "lexical involvement", "centrality", and "cohesiveness" (which are also values between 0 and 1). Although the sense of these components is somewhat similar to analogues implemented in the MetaMap, we compute them in the different way.

The "lexical involvement" assesses how the tokens of the variant correspond to the tokens of the thesaurus term. The component ignores order of tokens and considers terms as a bag of words. Russian is a free word order language and terms in text can have permutations of tokens. We also note that tokens in MeSHRUS terms are often in uncommon order. Thus, ignoring word order is quite justified. The lexical involvement is a weighted harmonic mean F_β of two values: the total weight of tokens of a variant among tokens of a term from the thesaurus ("concept coverage") and the total weight of tokens of a term from the thesaurus among tokens of a variant ("variant coverage"). Unlike the MetaMap, which convolves the concept and the variant coverages linearly, we used harmonic mean to combine them because it more strictly penalizes the component if any of the two weights has a small value. Total weights sum up from weights of individual tokens, which are not equal in these sums. They are altered according to heuristics to make matches of tokens of certain types be more significant than others. For example, matches of nouns and verbs are considered more significant than matches of adjectives and adverbs. It is possible to tune significance of "concept" and "variant coverage", the better performance is usually achieved when lexical involvement is inclined to the concept coverage.

The "centrality" shows whether the most significant token of a variant from syntax perspective (the head of the phrase) is present in the thesaurus term. It is 1 if the syntactic head of the biggest phrase in the variant is present in the term of the thesaurus and 0 otherwise. This component can prune variants with many less significant token matches that lack the most important match of the syntactic head.

The "cohesiveness" extends the idea of "lexical involvement" to syntactically connected phrases. It is a weighted harmonic mean of two values "coherence" and "variant coverage". The "coherence" is a maximum ratio of a number of syntactically connected tokens in the variant that participate in the match with the thesaurus term and a number of tokens in the variant. This component can prune variants with matches that are not syntactically linked and therefore are not related.

When the similarity is computed for all built variant - thesaurus term pairs, the result pairs are selected by a threshold. The selected variants are also filtered by some heuristics that prune nested variants with the same concept identifiers. The final variants are mapped to the semantic types according to their identifier. Variants that correspond SRD are mapped to "drugs". To map variants that correspond UMLS Metathesaurus terms we used UMLS Semantic Network. In this network, CUIs are mapped to semantic types and we map these types to several categories: diseases, symptoms, body sites.

3.2. Negation and "not patient" flag detection

It is crucial to know whether the found disease or symptom mentions are negated in text and whether they are associated with the patient or with another person, e. g. patient's relative, since many clinical notes also contain information about patient's heredity.

For negation detection, medical text processing systems usually apply simple approach based on pattern matching. For example, NegEx (Chapman et al., 2013) implements an algorithm that searchers for a list of patterns in a linear context in a window around a disease mention. Despite the simplicity of the algorithm, it proved itself robust with moderate performance and was adopted in cTAKES.

Besides the aforementioned strategy, we also implemented pattern search in a syntax tree, which showed somewhat better performance. The latter approach was integrated into the final pipeline. The developed negation detection module searches for the following patterns:

- particle "не" ("not") syntactically depends on one of the tokens of the disease/symptom;
- particle "не" ("does not") syntactically depends on predicate (single predicate word of complex predicate with auxiliary verb) that governs a token of the disease/symptom term;
- particle "нет"("no") governs a token from the disease/symptom term;
- a token of the disease/symptom term is governed by negation predicate, e. g. "отсутствует" ("is absent");
- particle "нет" ("no") immediately follows a disease/symptom mention.

Module that detects "not patient" flag mainly searches for mentions of patient relatives and patterns that associate them with disease terms in text:

- "у" + "relative mention" syntactically connected to or precedes disease term in a sentence;
- "наследственность" ("heredity") precedes disease mention in a sentence.

3.3. Discovering severity, course, and body site attributes of diseases

Severity attributes capture and normalize text cues that clarify whether the disease mentioned in text is "slight", "moderate", or "severe". Course attributes capture and

normalize text cues about disease progress: “worsened”, “changed”, “improved”, “resolved”. Body site attributes determine which body part disease mentions are associated with in a clinical text. In example: “*Диагностирована астма, атопическая, легкое персистирующее течение, ремиссия. … Имеется ангиопатия сетчатки.*” (“*Diagnosed asthma, atopic, mild persistent, during remission. … Retinal angiopathy is present.*”). The string “*легкое персистирующее течение*” (“*mild persistent”*) should be captured in the severity attribute of the disease mention “*астма, атопическая*” (“*asthma, atopic*”) and normalized as “slight”; the string “*ремиссия*” (“*during remission”*) should be captured in the course attribute of the same disease mention and normalized as “improved”; the string “*сетчатки*” (“retinal”) should be marked as body site and linked to the disease mention “*ангиопатия*” (“*angiopathy*”). These attributes no doubt are a very important piece of information that can be extracted from clinical narrative. The common practical usage of the extracted information is summarizing electronic health records for information or analytical systems in terms of standardized format like Clinical Document Architecture[10]. For the tasks of discovering severity, course, and body site attributes, we implemented several modules based on supervised machine learning methods.

The modules that discover severity and course attributes are almost the same. They consist of two separate submodules for the attribute span identification with linking it to the corresponding disease mention and for the attribute normalization. For the given disease mention, the submodule for the attribute span identification scans tokens of a sentence, in which the corresponding disease mention is located, and applies to them binary classifier that predicts whether token is an attribute related to the disease mention or not. When every token is classified, the tokens marked by the same attribute are grouped into continuous annotations. For severity and course span identification, we used the following lexical and syntactic features: lemmas and postags of tokens in a window around the classified token; whether the classified token syntactically depends on the disease term; distance between the classified token and the disease term; relative position of the token regarding to the given disease mention; number of disease annotations between the target disease mention and the token. When attribute spans are identified, another submodule with a separate classifier normalizes them. The feature set of the latter classifier is composed of token lemmas lying in the corresponding span of severity or course annotation represented as a bag of words.

In the task of discovering body site attributes, the spans that represent body parts are identified by thesaurus-based parser described in section 3.1. Therefore, only linking these spans to disease mentions is required. The module for linking body sites to diseases scans all body site—disease pairs within a sentence and analyzes them with a binary classifier. The features of the classifier include distance in tokens between a disease mention and a body site, whether they are syntactically linked, whether they are attached to the same word (e. g., predicate), the postag of this word, the number of disease mentions between the given disease mention and the body site.

We tested several classifiers for each of the tasks and subtasks: linear SVM, SVM with radial basis kernel, random forest, and AdaBoost. The parameter tuning was performed on the developed corpus via cross-validation.

---

[10] http://www.hl7.org/implement/standards/product_brief.cfm?product_id=7

4. Annotated corpus of clinical notes in Russian

In conjunction with specialists of Scientific Center of Children Health (SCCH)[11] we created annotated corpus of clinical free-text notes in Russian. The corpus is based on medical histories of more than 60 SCCH patients with allergic and pulmonary disorders and diseases. It comprises discharge summaries, radiology, echocardiography, ultrasound diagnostics reports, recommendations, and other records created by different physicians. The documents in the corpus were de-identified: all names were removed and dates were altered. With the help of SCCH experts, we developed an annotation scheme and a guideline. The scheme encompasses span annotations, their attributes, and relations. Table 1 describes the annotation scheme in detail.

Table 1. Annotation scheme

Annotation Type	Annotation Name	Description
Main annotations	Disease	The minimal span that is considered as a disease or disorder. Has attributes: Negation, NotPatient, Conditional.
	Symptom	The span that is considered as a symptom. It also can be a disorder. Has attributes: Negation, NotPatient, Conditional.
	Drug	The span that is a mention of a drug.
	Treatment	The span that contains the medical procedure aimed at disease treatment. Has attributes Effect, NotPatient.
	Body location	The span that is considered as a body site. Can be linked to Disease annotations.
	Severity	Span that captures severity of a disease. Always linked to Disease annotations. Has normalization value SType.
	Course	Span that captures course of a disease. Always linked to Disease annotations. It has normalization value CType.
Attributes	Negation	Boolean value associated with Disease or Symptom. Indicates that associated annotation is negated.
	NotPatient	Boolean value associated with Disease, Symptom, or Treatment. Indicates that associated annotation is not related to the patient (e. g. associated with a relative).
	Conditional	Boolean value associated with Disease or Symptom. Indicates that associated annotation appears under certain circumstances.
	Effect	Normalization value from the set {Positive; Negative; NoEffect} associated with Treatment. Summarizes the effect of a treatment.
	Degree	Normalization value from the set {Slight; Medium; Hard} for Severity annotation.
	CType	Normalization value from the set {changed; improved; worsened; resolved} for Course annotation.

[11] http://www.nczd.ru/eng/

The annotation scheme partially coincides with the CLEF eHealth 2014 Task 2. It lacks temporal information and some minor attributes. However, it is also in some way broader than CLEF eHealth annotation scheme because besides the annotations related to diseases it also includes annotations of drugs, treatments, and symptoms with their attributes.

The corpus was annotated and verified by physicians. For annotation purposes, we used Brat—the web-based tool, which was originally created for BioNLP challenge[12]. The Fig. 1 illustrates a fragment of an annotated text.

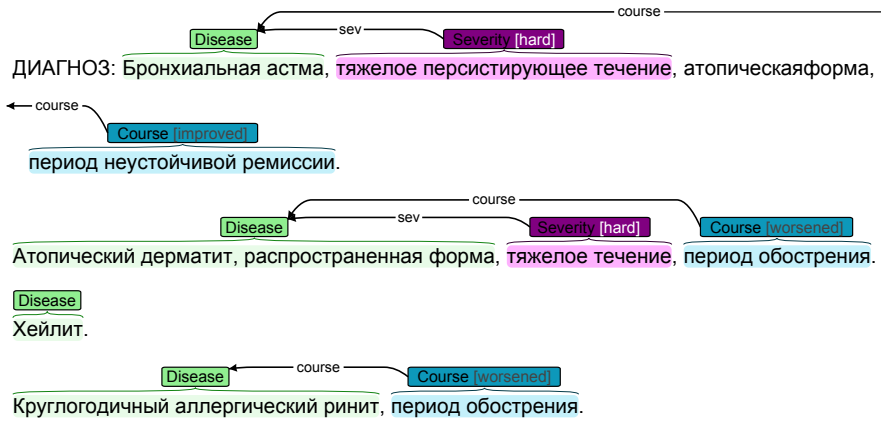

Fig. 1. Example of annotated text

The corpus currently contains more than 112 fully annotated texts with almost 45,000 tokens. There are more than 7,600 annotated entities and more than 4,000 annotated attributes and relations. The work on the corpus is still in progress: we are adding new texts and planning to expand the annotation scheme. The final goal is to make a verified gold standard for testing and resource for training machine-learning algorithms.

The corpus is freely available for the research community through the data use agreement[13]. However, the human subjects training certificate[14] would be required for access to it since the corpus contains the medical data of real patients. We are expecting this resource could be useful for creating and testing new applications and techniques for natural language processing of clinical texts in Russian.

---

[12] http://brat.nlplab.org/

[13] http://nlp.isa.ru/datasets/clinical

[14] https://phrp.nihtraining.com/users/login.php

5. Experiments

5.1. Disease and drug identification performance

The testing of medical term identification module was performed on a randomly selected holdout consisting of 30 texts; the rest of the corpus was used for parameter tuning. We calculated relaxed versions of precision, recall, and F_1 score. In the relaxed assessment, a span overlapping a gold standard span is considered correct.

Especially for the disease identification task, we prepared two baselines. The first baseline marks in text all words of thesaurus concepts related to "disease" semantic type. This baseline tends to maximum recall. The second baseline marks in text only token chains that exactly match a whole bag of words of a thesaurus concept. This baseline tends to maximum precision. Table 2 presents performance of the developed module and the baselines.

Table 2. Performance of the disease identification module and the baselines

Module	Recall,%	Precision,%	F_1-score,%
Disease identification	72.8	95.1	**82.4**
Baseline 1	**84.9**	9.3	16.7
Baseline 2	69.8	**99.2**	81.9

As expected, the implemented module has better overall performance than the baselines. However, the performance of the baseline 2 appeared to be very close to the performance of the developed module. This is because we used relaxed measures to match the gold standard with the answers of the system. Many complex medical terms in text include single words that represent in thesauri general disease categories and marking these words is also considered as a match. The developed method should much more significantly outmatch the baseline 2 in the task of mapping terms to concepts because the baseline 2 often maps terms to the very general concepts although they should be mapped to ones that are more specific. For example, the developed module finds term "аллергический ринит" ("allergic rhinitis") in text. This string cannot be mapped exactly to the bag of words of any thesaurus term, so the baseline 2 will not find it. However, it will find "ринит" ("rhinitis") and map it to the more general concept. Both answers will be considered as a match for the disease identification task, however, in the concept-mapping task, the answer of the developed module will be the only correct one. For now, we cannot evaluate concept-mapping task since we did not annotate concept identifiers in the test set.

For the evaluation of the drug identification performance, we used exactly the same framework as for the disease identification. The precision was 84.3, the recall was 74.6, and the F1-score was 79.2. The obtained results illustrate that the SRD is the suitable source for the drug identification task. The recall is somewhat lower than expected because annotators marked not only registered drugs but also mentions of therapeutic cosmetics (e.g., anti-allergic cream). We also note that some false negatives were due to corpus texts contained contractions and general names of drugs (e.g., "пенициллин" ("penicillin") is not present in SRD but "бензилпенициллин" ("benzathine penicillin") is in SRD).

5.2. Symptom negation and "not patient" flag detection performance

We found that the annotated corpus contains just a few negations of disease mentions, much more negations are related to symptoms. Therefore, instead of disease negation detection we only evaluated the performance of the symptom negation detection.

The number of negation and "not patient" flag annotations in corpus is relatively small—both around 100. To make the test set representative, we had to evaluate negation and "not patient" flag detection using the whole corpus rather than the selected holdout. Since the developed modules rely primarily on rules and do not have many parameters, such approach has a minor effect on the performance assessment bias.

In the evaluation, we only took into account negation and "not subject" flag annotations that are attributes of symptom and disease mentions identified by our system to exclude false negatives of the term identification module.

We calculated the precision, recall and F_1-score. The table 3 presents evaluation results of the symptom negation and "not subject" detection.

Table 3. Performance of the symptom negation and "not subject" detection

Module	Recall,%	Precision,%	F_1-score,%
Symptom negation	98.7	95.3	97.0
Disease "not patient"	90.9	96.8	93.8

The obtained results show that rather simple approach and few patterns can cover the most cases of negation and "not patient" attributes in clinical texts. However, we admit that the current test set is small and not very representative for the perfect evaluation.

5.3. Performance of discovering severity, course, and body site attributes of diseases

For all evaluations and for tuning classifier parameters of the modules, we used 5-fold cross validation on the annotated corpus.

We separately evaluated performance of the identification and normalization of severity and course attributes. To exclude the false negatives of the disease identification module, we only took into account course and severity annotations that are related to disease mentions identified by our system. For evaluation of severity and course identification, similarly to the evaluation of medical term identification, we calculated relaxed versions of precision, recall, and F_1-score. Tables 4 and 5 present the performance of course and severity identification respectively.

Attribute normalization is a multilabel classification task without "empty" class. Therefore, evaluation of severity and course normalization was performed via accuracy. Table 6 presents the performance of severity and course normalization.

The performance of linking body sites to diseases was evaluated via relaxed precision, recall, and F_1-score. Only disease mentions that were identified by the system were taken into account. Table 7 shows the results of the evaluation.

The table in Appendix 1 contains examples of annotations in the corpus and annotations generated by the final pipeline that uses the best classifiers.

Table 4. Performance of severity identification

Classifier	Recall,%	Precision,%	F_1-score,%
Linear SVM	99.2	41.7	58.6
RBF SVM	95.0	80.8	87.1
Random forest	93.6	82.6	**87.5**
AdaBoost (Dec. tree)	97.3	75.2	84.7

Table 5. Performance of course identification

Classifier	Recall,%	Precision,%	F_1-score,%
Linear SVM	92.3	99.2	**95.7**
RBF SVM	88.3	99.3	93.4
Random forest	88.3	99.3	93.4
AdaBoost (Dec. tree)	90.0	98.4	93.9

Table 6. Performance of severity and course normalization

Module	Classifier	Accuracy,%
Severity normalization	Linear SVM	88.4
	RBF SVM	88.0
	Random forest	89.3
	AdaBoost (Dec. tree)	**89.8**
Course normalization	Linear SVM	89.4
	RBF SVM	91.4
	Random forest	**92.7**
	AdaBoost (Dec. tree)	91.4

Table 7. Performance of linking body sites to diseases

Classifier	Precision, %	Recall, %	F_1-score, %
Linear SVM	85.4	77.5	81.0
RBF SVM	91.4	76.6	**83.3**
Random forest	86.6	75.8	80.8
AdaBoost (Dec. tree)	84.0	76.6	79.9

The obtained results illustrate that the task of discovering attributes of disease mentions in clinical texts in Russian can be successfully solved by supervised machine learning techniques. Experiments with different classifiers showed statistical difference between their results only in the task of severity identification. This can be because of importance of word collocations that are differently modeled by these classifiers. The developed modules represent a solid baseline. However, there is still a big space for improvement, which can be achieved by developing richer set of features and applying state-of-the-art machine learning methods, e. g. conditional random fields. We consider that the obtained results can be a useful landmark for the future research.

6. Conclusion and future work

We presented the pipeline for processing of clinical texts in Russian and the corpus of annotated clinical notes. We evaluated the pipeline and showed that it can successfully solve the key tasks of information extraction from clinical narrative. In the ongoing work, we are expanding annotated corpus to make it more representative and suitable for training machine-learning algorithms, as well as for reliable testing of clinical text processing methods and tools. In the future work, we are planning to upgrade corpus annotation scheme and include in the pipeline modules for discovering treatments with modifiers and temporal expressions related to disease mentions. We also are planning to apply the developed pipeline for the high-level task of clinical information retrieval and clinical data analysis.

Acknowledgments

We are grateful to experts of Scientific Center of Children Health for help on annotating corpus of clinical texts in Russian. This work was supported by RFBR, project 13-04-12062.

References

1. *Apresjan J. D., Boguslavskij I. M., Iomdin B. L., Iomdin L. L., Sannikov A. V., Sannikov V. G. and Sizov L. L.* (2005), Syntactically and semantically annotated corpus of Russian language: Present state and perspectives [Sintaksicheski i semanticheski annotirovannyj korpus russkogo jazyka: sovremennoe sostojanie i perspektivy], National Corpus of Russian Language: 2003–2005 [Natsional'nyj korpus russkogo jazyka: 2003–2005], pp. 193–214, (in Russian)
2. *Aronson, A. R. and Lang, F.-M.* (2010), An overview of MetaMap: historical perspective and recent advances, Journal of the American Medical Informatics Association (JAMIA), (3), Vol. 17, pp. 229–236
3. *Bos, L. and Donnelly, K.* (2006), SNOMED-CT: The advanced terminology and coding system for eHealth, Studies in health technology and informatics, Vol. 121, pp. 279–290

4. *Chapman, W. W., Hilert, D., Velupillai, S., Kvist, M., Skeppstedt, M., Chapman, B. E., Conway, M., Tharp, M., Mowery, D. L. and Deleger, L.* (2013), Extending the NegEx lexicon for multiple languages, Studies in health technology and informatics, Vol. 192, pp. 677–681
5. *Dligach, D., Bethard, S., Becker, L., Miller, T. A. and Savova, G. K.* (2014), Discovering body site and severity modifiers in clinical texts, Journal of the American Medical Informatics Association (JAMIA), pp. 448–454
6. *Hamon, T., Grouin, C. and Zweigenbaum, P.* (2014), Disease and Disorder Template Filling using Rule-based and Statistical Approaches, In CLEF (Working Notes), pp. 79–90
7. *Huynh, H. N. and Ho, S. L. V. B. Q.* (2014), ShARe/CLEFeHealth: A Hybrid Approach for Task 2, In CLEF (Working Notes), pp. 103–110
8. *Johri, N., Niwa, Y. and Chikka, V. R.* (2014), Optimizing Apache cTAKES for Disease/Disorder Template Filling: Team HITACHI in the ShARe/CLEF 2014 eHealth Evaluation Lab, In CLEF (Working Notes), pp. 111–123
9. *Lipscomb, C. E.* (2000), Medical subject headings (MeSH), Bulletin of the Medical Library Association, Vol. 88, pp. 265–266
10. *Liu, Y.-C. and Ku, L.-W.* (2014), CLEFeHealth 2014 Normalization of Information Extraction Challenge using Multi-model Method, In CLEF (Working Notes), pp. 124–132
11. *McCray, A. T.* (1989), The UMLS Semantic Network. In Proceedings of the Annual Symposium on Computer Application in Medical Care, pp. 503–507, American Medical Informatics Association
12. *Mkrtchyan, T. and Sonntag, D.* (2014), Deep Parsing at the CLEF2014 IE Task, In CLEF (Working Notes), pp. 138–146
13. *Névéol, A., Grouin, C., Leixa, J., Rosset, S. and Zweigenbaum, P.* (2014), The Quaero French medical corpus: A resource for medical entity recognition and normalization, In Proceedings of LREC BioTxtM 2014 Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing
14. *Nivre, J., Boguslavsky, I. M. and Iomdin, L. L.* (2008), Parsing the SynTagRus treebank of Russian, In Proceedings of the 22nd International Conference on Computational Linguistics, pp. 641–648, Association for Computational Linguistics
15. *Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S. and Marsi, E.* (2007), MaltParser: A language-independent system for data-driven dependency parsing, Natural Language Engineering, (2), Vol. 13, pp. 95–135
16. *Pradhan, S., Elhadad, N., South, B. R., Martinez, D., Christensen, L., Vogel, A., Suominen, H., Chapman, W. W. and Savova, G.* (2015), Evaluating the state of the art in disorder recognition and normalization of the clinical narrative, Journal of the American Medical Informatics Association (JAMIA), (1), Vol. 22, pp. 143–154
17. *Ramanan, S. V. and Nathan, P. S.* (2014), Cocoa: Extending a Rule-based System to Tag Disease Attributes in Clinical Records, In CLEF (Working Notes), pp. 150–155
18. *Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C. and Chute, C. G.* (2010), Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications, Journal of the American Medical Informatics Association (JAMIA), (5), Vol. 17, pp. 507–513

19. *Schuyler, P. L., Hole, W. T., Tuttle, M. S. and Sherertz, D. D.* (1993), The UMLS Metathesaurus: Representing different views of biomedical concepts, Bulletin of the Medical Library Association, (2), Vol. 81, 217–222
20. *Sequeira, J., Miranda, N., Goncalves, T. and Quaresma, P.* (2014), TeamUEvora at CLEF eHealth 2014 Task2a, In CLEF (Working Notes), pp. 156–166
21. *Sharoff S. and Nivre J.* (2011), The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011", pp. 591–604
22. *Smirnov, I. V., Shelmanov, A. O., Kuznetsova, E. S. and Hramoin, I. V.* (2014), Semantic-syntactic analysis of natural languages. Part II. Method for semantic-syntactic analysis of texts [Semantiko-sintaksicheskij analiz estestvennyh jazykov Chast' II. Metod semantiko-sintaksicheskogo analiza tekstov], Artificial Intelligence and Decision Making [Iskusstvennyj intellekt i prinjatie reshenij], (1), pp. 95–108, (in Russian)
23. *Sokirko, A.* (2001), A short description of Dialing Project, available at: http://www.aot.ru/
24. *Sun, W., Rumshisky, A. and Uzuner, O.* (2013), Evaluating temporal relations in clinical text: 2012 i2b2 challenge, Journal of the American Medical Informatics Association (JAMIA), (5), Vol. 20, pp. 806–813
25. *Suominen, H., Salanterä, S., Velupillai, S., Chapman, W. W., Savova, G., Elhadad, N., Pradhan, S., South, B. R., Mowery, D. L., Jones, G. J. et al.* (2013), Overview of the ShARe/CLEF eHealth evaluation lab 2013, In Information Access Evaluation. Multilinguality, Multimodality, and Visualization, Lecture Notes in Computer Science, Vol. 8138, pp. 212–231, Springer.

Appendix 1. Example of corpus and parser annotations

Annotation	Corpus	Parser
ДИАГНОЗ: Бронхиальная астма, тяжелое персистирующее течение, атопическая форма, период неустойчивой ремиссии. Атопический дерматит, распространенная форма, тяжелое течение, период обострения. Хейлит. Круглогодичный аллергический ринит, период обострения. … Пролапс митрального клапана с регургитацией 3-4 мм. Грудной сколиоз 1 степени. Вегетососудистая дисфункция по гипотоническому типу.		
Disease	Бронхиальная астма	Бронхиальная астма (C0004096)
Severity	тяжелое персистирующее течение (hard)	тяжелое персистирующее течение (hard)
Course	период неустойчивой ремиссии (improved)	период неустойчивой ремиссии (improved)
Disease	Атопический дерматит распространенная форма	дерматит (C0011603) | Атопический дерматит (C0011615)
Severity	тяжелое течение (hard)	Течение (medium)
Course	период обострения (worsened)	период обострения (worsened)
Disease	Круглогодичный аллергический ринит	ринит (C0035455) | аллергический ринит (C0018621, C0035457)
Course	период обострения (worsened)	период обострения (worsened)
Disease	Пролапс митрального клапана с регургитацией 3–4 мм	Пролапс (C0033377) | Пролапс митрального клапана (C0026267, C0003505, C0040962, C0079485)
Body location	митрального клапана	митрального клапана
Disease	Грудной сколиоз	Сколиоз (C0036439)
Severity	1 степени (light)	1 степени (light)
Консультация ЛОР врача. Катаральный риносинусит. Назначения согласованы. Проводилось лечение: сумамед 10 мг/кг, тридерм местно, ксимелин 2–3 к. × 3 раза в день, креон по 1 к. × 3 раза в день, полидекса по 1 инст. × 3 раза в день. Выписывается домой в удовлетворительном состоянии под наблюдение педиатра, аллерголога. Контактов с инфекционными больными не было.		
Disease	Катаральный риносинусит	
Drug	сумамед	сумамед
Drug	тридерм	тридерм
Drug	ксимелин	ксимелин
Drug	креон	креон
Drug	полидекса	полидекса
Наследственность – у матери лекарственная аллергия на пенициллины, у старшего брата пищевая аллергия, у прабабушки по о/л бронхиальная астма. Элиминационный режим не соблюдается: дома рыбки, кошка		
Disease	лекарственная аллергия (NotPatient = True)	Аллергия (C0020517) | лекарственная аллергия (C0020517) (NotPatient = True)
Drug	пенициллины	
Disease	пищевая аллергия	Аллергия (C0020517) | пищевая аллергия (C0020517) (NotPatient = True)
Disease	бронхиальная астма	бронхиальная астма