Information extraction from clinical texts in Russian

Shelmanov A. O.*, Smirnov I. V.*, Vishneva E. A.**

shelmanov@isa.ru

* Institute for systems analysis of Russian academy of sciences, Moscow, Russia

** Scientific Centre of Children Health, Moscow, Russia

2015, Moscow

Outline

- Natural language processing in medicine and its applications
- The task
- Related work
- Corpus of clinical texts in Russian
- The pipeline
- Methods
- Evaluation
- Conclusion

Natural language processing in medicine

Processing of biomedical literature

Mainly tools that support information retrieval from scientific publications for biomedical research

Several well-known tasks:

- NER: genes, proteins, diseases, drugs
- Mining disease-gene associations
- Extraction of drug-drug interactions

Clinical text processing

Tools that support information retrieval for health care and research on clinical data

Clinical text processing applications

- Many clinical records are free text:

  radiology, echocardiography, and electrocardiogram reports, anamnesis, results of ultrasound diagnostics, discharge summaries

- NLP can be used for:
  - Patient medical history management
  - Decision support systems in medical domain
  - Research on clinical data written in free text
  - Converting health records into patient-friendly form to help patients understand it
  - Structuring health records for interoperability between healthcare providers (e.g. “Clinical Document Architecture” standard)
  - Coding health records for exchange between healthcare providers and insurance companies (ICD-10 codes)

Peculiarities of clinical text processing

Many common NLP systems are trained to process texts like news articles, tweets, etc.

Clinical narrative is very specific:

- Terse and compressed
- Medical lexis: medical, biological terms
- Abbreviations
- Mistakes

Therefore, peculiarities:

- Need specific annotated corpora
- Need specific thesauri and ontologies
- Specific tasks due to specific goals and peculiarities of clinical texts

The task

- Developing information-analytical system for Scientific Centre of Children Health (SCCH)
- Need pipeline for information extraction from clinical texts in Russian
- Tasks of information extraction:
  - Finding medical terms: **diseases**, **body locations**, **drugs**, treatments, …
  - Normalizing terms: mapping terms in text to concepts in medical thesaurus
  - Discovering attributes of terms (currently only diseases and symptoms): Severity, Course, Body locations, Negation, “Whether disease related to patient or not” = NotPatient flag detection
  - Discovering = identification + normalization
- Development and evaluation requires annotated corpus

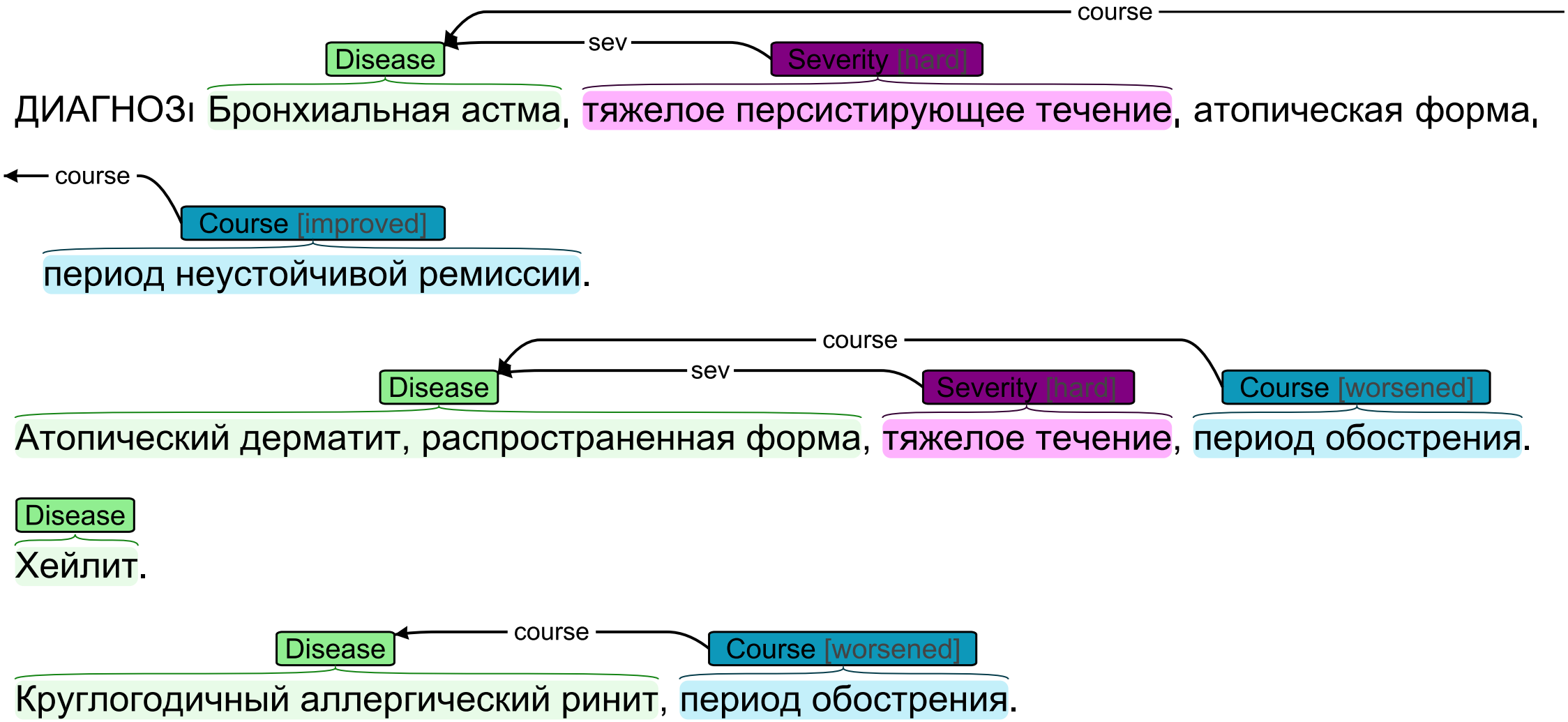
Related work

- NLP for medicine is a rapidly developing research area
- Many shared tasks and workshops, e.g.:
  - CLEF eHEALTH (2013 – 2015)
  - SemEval (2014, 2015)
  - i2b2 (2008 – 2014)
  - BioNLP-ST (2009 – 2015)
- Many NLP systems, e.g.:
  - MedLEE (Friedman C., 2000)
  - HiTEX (Qing T Zeng et al., 2006)
  - cTAKES (Mayo Clinic) (Savova et al., 2010)
- Other related work:
  - Body site and severity extraction (Dligach D. et al., 2014)
  - Disorder identification and normalization (Pradhan S., 2014)

Corpus of clinical texts in Russian (1)

- Corpus annotated by specialists of Scientific Center of Children Health (SCCH)
- Corpus consists of 60 medical histories of SCCH patients with allergic and pulmonary disorders and diseases
- Comprises discharge summaries, radiology, echocardiography, ultrasound diagnostics reports, recommendations
- Annotations: “Disease”, “Symptom”, “Drug”, “Treatment”, “Effect” of treatment, “Body location”, “Severity”, “Course”, “Negation”, “NotPatient” + normalization values and relations

Annotation example



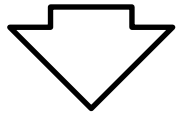
Annotated with Brat – the web-based tool http://brat.nlplab.org/

Corpus of clinical texts in Russian (2)

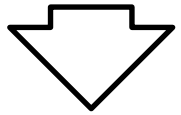
- Almost 45,000 tokens. More than 7,600 annotated entities and more than 4,000 annotated attributes and relations
- Work is in progress
- English related resources:
  - ShARe (Mowery D. L., 2014)
  - SHARPn (Pradhan S., 2015)
- Anonymized: removed names, altered dates
- Available for research purposes at http://nlp.isa.ru/datasets/clinical
- Requires human subjects training certificate!

The pipeline for information extraction from clinical texts in Russian

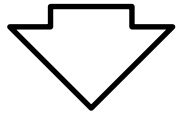
Tokenizing, splitting, PosTagging by AOT.ru (Sokirko, 2001)



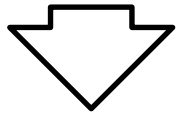
Parsing by MaltParser trained on SynTagRus (Apresjan et al., 2005)



Medical term identification and normalization



Negation and “not patient” flag detection



Discovering severity, course, and
body site attributes of diseases

Identifying medical terms in text and mapping them to thesauri (1)

Developed method adopts MetaMap approach (Aronson et al., 2010), linguistically motivated tool for mapping terms from medical texts to concepts in UMLS Metathesaurus:

- Generate extensive amount of term variants from text expressions
- Perform fuzzy comparison between the variants and the thesaurus terms
- Rank the variants by heuristically reasoned score
- Pick the most confident ones

Thesauri for processing clinical texts in Russian

- UMLS Metathesaurus + UMLS Semantic network
  - Metathesaurus maps concepts disjoint medical thesauri to the CUIs (concept unique identifier)
  - Incorporates MeSH, SNOMED-CT, ICD-10, etc.
  - The only resource in Russian is MeSHRUS ~ 27,000 concepts; 85,000 terms
  - Semantic network maps CUIs to semantic types like organisms, chemicals, diseases, symptoms etc.
  - Using these resources for disease, symptom and body location identification
- State Register of Drugs
  - Database of all drugs officially registered and allowed for sale in Russia
  - We grouped drugs with similar active chemicals into concepts of thesaurus
  - Got 3,600 unique concepts and almost 12,000 terms
  - Using the resource for identification of drugs in clinical texts

Identifying medical terms in text and mapping them to thesauri (2)

- Find keywords from thesaurus terms using its inverted index
- Generate variants for keywords using: syntax relations, linear context around the keyword
- Compare variants with terms in thesauri:
  - lexical involvement – weighted harmonic mean (F_β) of:
    - the total weight of tokens of a variant among tokens of a term from the thesaurus
    - the total weight of tokens of a term from the thesaurus among tokens of a variant
  - Centrality:
    - 1 if the syntactic head of the biggest phrase in the variant is present in the term of the thesaurus
    - 0 otherwise
  - Cohesiveness – extends the idea of “lexical involvement” to syntactically connected phrases
- Select variant - thesaurus term pairs by threshold and heuristics

Detection of disease/symptom negation

- We implemented rule-based approach
- The rule-based approach is widely used, e.g., NegEx (Chapman W. W., 2013) algorithm searches for simple patterns in a linear context in a window around a disease mention
- Implemented pattern search in a syntax tree
- Patterns for negation detection:
  - “не” (“not”) syntactically depends on one of the tokens of the disease/symptom
  - particle “не” (“does not”) syntactically depends on predicate
  - 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

NotPatient flag detection

- Rule-based approach
- Searching for mentions of relatives in sentence because many clinical notes describe heredity
- Patterns for NotPatient flag detection:
  - “у” (“has”) + “relative mention” syntactically connected to or precedes disease term in a sentence
  - “наследственность” (“heredity”) precedes disease mention in a sentence

Extraction of severity and course

- Machine learning on developed corpus, applied: linear SVM, rbf SVM, random forest, AdaBoost
- Two separate submodules:
  - Severity span identification for corresponding disease mentions
  - Normalization
- Classify tokens one by one and predict whether it is a part of severity/course annotation, linked to the given disease annotation
- Lexical and syntactic features: lemmas and postags of tokens in a window around the classified token; whether the classified token syntactically depends from 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 disease mention and the token

Normalization of severity and course

- Using the same machine learning techniques
- Classifying spans identified as severity/course annotations
- Feature set consists of token lemmas lying in the corresponding span of severity/course annotation represented as a bag of words

Linking body sites to the disease mentions

- Body locations and diseases are identified by thesaurus-based method
- Predicting relation between identified annotations for body locations and diseases
- Using same machine learning techniques
- Features set: 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

Evaluation of disease identification

- For disease identification created two baselines:
  - Baseline 1 marks in text all words of thesaurus concepts related to “disease” semantic type (maximum recall)
  - Baseline 2 marks in text only token chains that exactly match a whole bag of words of a thesaurus term (maximum precision)
- Calculated relaxed precision, recall, F1

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

Evaluation of drug identification

- The same evaluation framework as for disease identification
- Results:
  - Precision = 84.3
  - Recall = 74.6
  - F_1-score = 79.2
- SRD is “ok” for finding drugs
- Mistakes:
  - Annotators marked not only registered drugs but also mentions of therapeutic cosmetics
  - “Пенициллин” (“penicillin”) not present in SRD, but “бензилпенициллин” (“benzathine penicillin”)
  - Some normalization problems

Evaluation of disease and symptom negation detection

- The number of negations for diseases is small
- Evaluated negation detection both for disease and symptom annotations

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

- Rather simple patterns can handle this task
- Corpus is not very representative for evaluation (about 100 samples)

Evaluation of course and severity identification

- Used 5-cross validation on the annotated corpus for all tasks
- Calculated relaxed precision, recall, F1
- Tested different classifiers
- 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

- 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

Evaluation 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

Evaluation of the linking body locations 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

Comparing results with corpus

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)

Final remarks and future work

- Results:
  - Created and evaluated the pipeline for information extraction from clinical texts in Russian
  - Created annotated corpus
- Future work:
  - Extending corpus: the annotation scheme and the corpus size
  - Creating more tools for information extraction (treatments)
  - Apply the developed pipeline for the high-level task of clinical information retrieval and clinical data analysis

Thank you for your attention
any questions ?

References (1)

- Friedman C. A broad-coverage natural language processing system //Proceedings of the AMIA Symposium. – American Medical Informatics Association, 2000.
- Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evalua-tion of a natural language processing system / Qing T Zeng, Sergey Goryachev, Scott Weiss et al. // BMC medical informatics and decision making. — 2006. — Vol. 6, no. 30.
- 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
- 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
- 2001Sokirko, A. (2001), A short description of Dialing Project, available at: http://www.aot.ru/

References (2)

- 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)
- 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
- 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
- Mowery D. L. et al. Task 2: Share/clef ehealth evaluation lab 2014 //Proceedings of CLEF 2014. – 2014.