Kazan Federal UNIVERSITY

Laboratory of Big Data and Textual Analysis

Higher Institute of Information Technologies and Information Systems

28 April, 2017

Outline

- Laboratory of Big Data and Textual Analysis
- Past projects
- Current projects

Laboratory of Big Data and Textual Analysis

- Team:
    - Prof. V. Solovyev;
    - Postdocs: V. Ivanov, E. Tutubalina;
    - A. Sirotkin (HSE St. Petersburg), A. Kadurin (PDMI RAS, Mail.Ru Moscow);
    - 3 students (Master or PhD);
- Collaborations:
    - A. Tropsha, A. Kotov (USA);
    - I. Batyrshin (Mexico);
    - KFU Cheminformatics Lab, N. Loukachevitch, V. Polyakov, S. Nikolenko (Russia).

- RFBR, 13-07-00773a, The technology of development of systems for information extraction from Russian language texts, 2013-2015
- RFBR-Tatarstan Republic, 13-06-97065, Quantitative characteristics and comparison of evolution schemes for Turkic languages (using computer databases on grammar, lexis, genes), 2013-2014
- Hewlett-Packard, Agreement CW307214, project InfEx-2013, 2013, project EvExRus-2012, 2012
- “Dynasty Foundation”, DP-97/13, Development and realisation of hybrid technology of Information Extraction for Russian language, 2013-2014
- The Ministry of education and science of the RF, 02.740.11.0595, Languages of the World: information resources for typological, comparative and areal researches, 2010-2012
- The Ministry of education and science of the RF, 34.5517.2017/ВУ, State assignment "Organisation of scientific research ", 2014-2016, 2017-2019
- RSF, 15-11-10019, Text mining models and methods for analysis of the needs, preferences and consumer behavior, 2015-2017
- RFBR, 16-06-00165, Cognitive model for verbal inflectional paradigm in Russian: quantitative analysis of word frequency dynamics, 2016-2018
- RFBR, 15-29-01173, Computational models and mathematical methods for big data analysis of trends and correlations in society, 2015-2017
- RFBR, № 16-37-50070, Using deep learning models for problem of detection of social significant content, 2016-2016
- Russian Government Program of Competitive Growth of Kazan Federal University, 01-06/655, Medical informatics, 2016-2017.

Conferences

- International Conference on Web Search and Data Mining (WSDM 2016; USA)
- European Conference on Information Retrieval (ECIR 2015; Austria)
- International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013; Greece)
- International Conference on Computational Linguistics and Intellectual Technologies "Dialogue" (Dialogue 2015, 2016, 2017; Russia)
- International Conference on Text, Speech, and Dialogue (TSD 2014, 2015, 2016; Czech Republic)
- Mexican International Conference on Artificial Intelligence (2013, 2014, 2015; Mexico)
- Women in Machine Learning Workshop (WiML 2015 in conj. with NIPS, Canada)
- The 8th IEEE International Conference on Social Computing and Networking (SocialCom 2015; China)
- International Conference on Knowledge Engineering and Semantic Web (KESW 2013, 2016; Russia, Czech Republic)

Past projects

1. Development of Ontologies
    a. Ontology CIDOC CRM - RFBR / Kunstkamera / European Commission (europeana.eu)
    b. OntoMathPro ontology
    c. Nanomaterials Ontology - Ministry of Education and Science of the Russian Federation
2. Development of text mining methods for Russian
    a. NLP@CLOUD - Ministry of Education and Science of the Russian Federation
    b. Information Extraction - joint work with Hewlett-Packard (EvExRus-2012, InfEx-2013)
3. Development of linguistic resources for Russian
    a. Corpora: 18th Century Textual Documents in Russian National Corpus - joint work with Russian Language Institute of the Russian Academy of Sciences and Yandex
    b. Entities Dictionaries (e.g, Persons, Organizations)
    c. Thesaurus: Russian WordNet - joint work with Research Computing Center of Moscow State University

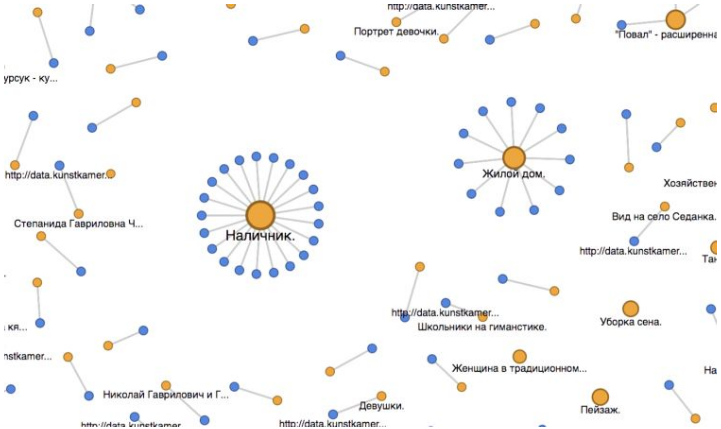
Open Kunstkammer Data Project

Goal: Represent a dataset consisting of more than 40 000 digital images and their descriptions in English and Russian as Linked Data

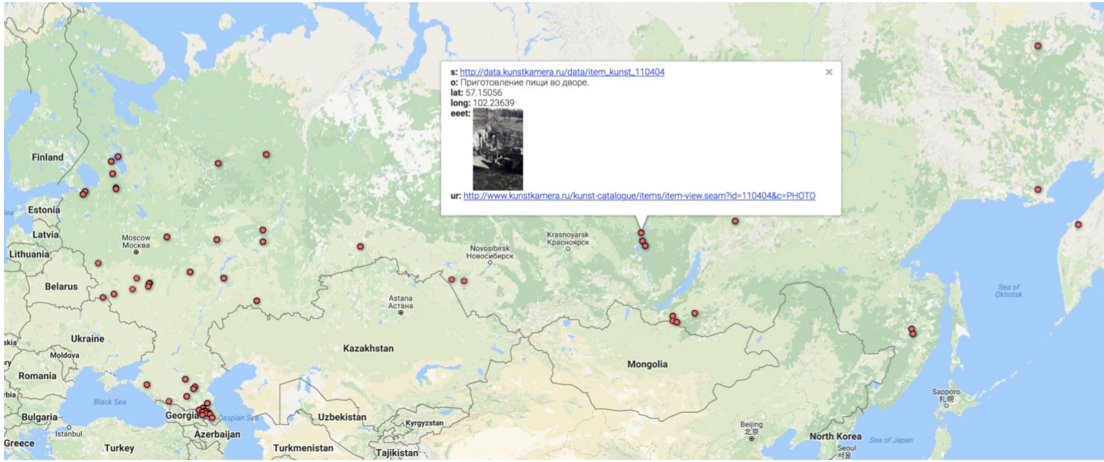
Constrains: Follow ICOM-CIDOC recommendations; “SKOSify” MAE RAS’s controlled bilingual vocabularies; interlink with **DBPedia and Geonames**

The resulting data warehouse (more than 10 M triples): interconnected descriptions of **museum objects, persons, places** and **events**

SPARQL end-point: http://data.kunstkamera.ru/sparql



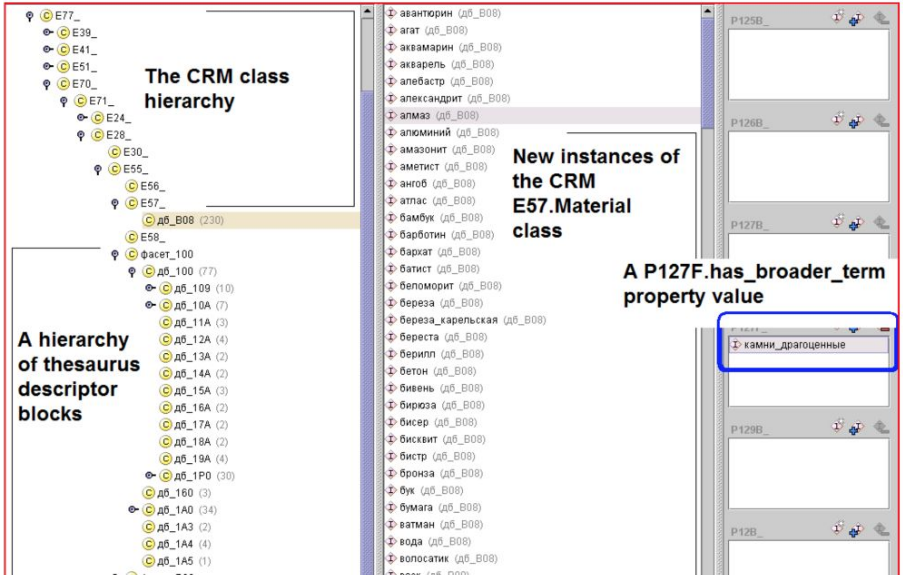
http://data.kunstkamer...
Портрет девочки.
"Повал" - расширенна
урсук - ку...
http://data.kunstkamer...
Наличник.
Жилой дом.
Хозяйствен
Степанида Гавриловна Ч...
Вид на село Седанка.
http://data.kunstkamer...
Та
http://data.kunstkamer...
Школьники на гимнастике.
Уборка сена.
кя...
nstkamer...
Женщина в традиционном...
На
Николай Гаврилович и Г...
Девушки.
Пейзаж.
http://data.kunstkamer...
http://data.kunstkamer...



s: http://data.kunstkamera.ru/data/item_kunst_110404
o: Приготовление пищи во дворе.
lat: 57.15056
long: 102.23639
eeet:
ur: http://www.kunstkamera.ru/kunst-catalogue/items/item-view.seam?id=110404&c=PHOTO

CIDOC Conceptual Reference Model and the Europeana

1) The **CIDOC CRM** ontology is a reference model in museum domain; it was translated into Russian
2) Then it was mapped to major thesauri in cultural heritage, including the Getty’s Art and Architecture Thesaurus **(more than 30,000 concepts)**
3) Museum metadata for KFU museum was represented in the CIDOC CRM format
4) Then the metadata was transferred to the European CH project: www.europeana.eu



The CRM class hierarchy

A hierarchy of thesaurus descriptor blocks

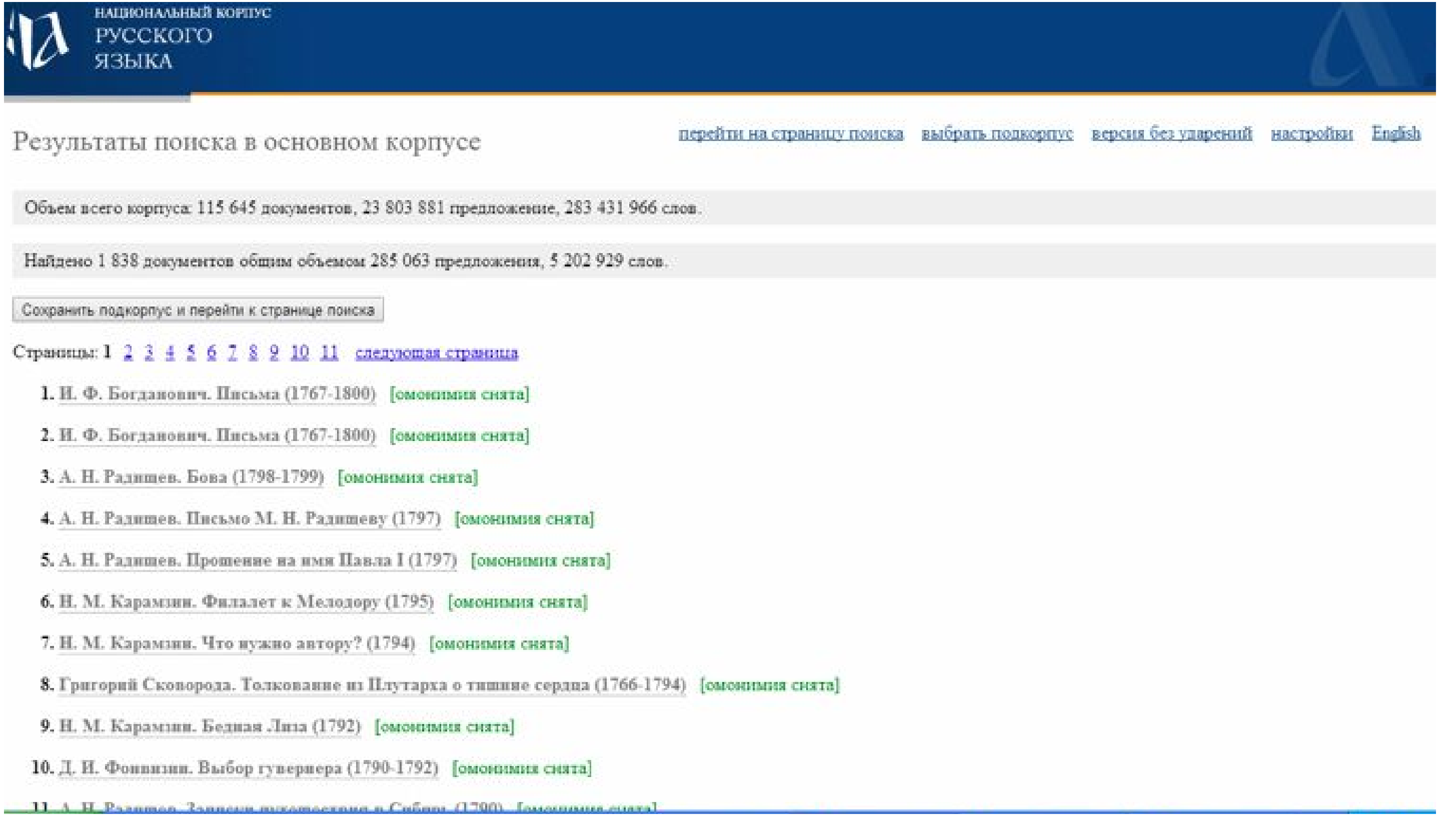
New instances of the CRM E57.Material class

A P127F.has_broader_term property value

```
E77_
  E39_
  E41_
  E51_
  E70_
    E71_
      E24_
      E28_
        E30_
        E55_
          E56_
          E57_
            дб_B08 (230)
          E58_
          фасет_100
            дб_100 (77)
              дб_109 (10)
              дб_10A (7)
                дб_11A (3)
                дб_12A (4)
                дб_13A (2)
                дб_14A (2)
                дб_15A (3)
                дб_16A (2)
                дб_17A (2)
                дб_18A (2)
                дб_19A (4)
              дб_1P0 (30)
            дб_160 (3)
            дб_1A0 (34)
            дб_1A3 (2)
            дб_1A4 (4)
            дб_1A5 (1)
```

```
авантюрин (дб_B08)
агат (дб_B08)
аквамарин (дб_B08)
акварель (дб_B08)
алебастр (дб_B08)
александрит (дб_B08)
алмаз (дб_B08)
алюминий (дб_B08)
амазонит (дб_B08)
аметист (дб_B08)
ангоб (дб_B08)
атлас (дб_B08)
бамбук (дб_B08)
барботин (дб_B08)
бархат (дб_B08)
батист (дб_B08)
беломорит (дб_B08)
береза (дб_B08)
береза_карельская (дб_B08)
береста (дб_B08)
берилл (дб_B08)
бетон (дб_B08)
бивень (дб_B08)
бирюза (дб_B08)
бисер (дб_B08)
бисквит (дб_B08)
бистр (дб_B08)
бронза (дб_B08)
бук (дб_B08)
бумага (дб_B08)
ватман (дб_B08)
вода (дб_B08)
волосатик (дб_B08)
```

18th Century Texts in RNC



НАЦИОНАЛЬНЫЙ КОРПУС
РУССКОГО
ЯЗЫКА

Результаты поиска в основном корпусе

перейти на страницу поиска
выбрать подкорпус
версия без ударений
настройки
English

Объем всего корпуса: 115 645 документов, 23 803 881 предложение, 283 431 966 слов.

Найдено 1 838 документов общим объемом 285 063 предложения, 5 202 929 слов.

Сохранить подкорпус и перейти к странице поиска

Страницы: 1 2 3 4 5 6 7 8 9 10 11 следующая страница

1. И. Ф. Богданович. Письма (1767-1800) [омонимия снята]
2. И. Ф. Богданович. Письма (1767-1800) [омонимия снята]
3. А. Н. Радищев. Бова (1798-1799) [омонимия снята]
4. А. Н. Радищев. Письмо М. Н. Радищеву (1797) [омонимия снята]
5. А. Н. Радищев. Прошение на имя Павла I (1797) [омонимия снята]
6. Н. М. Карамзин. Филалет к Мелодору (1795) [омонимия снята]
7. Н. М. Карамзин. Что нужно автору? (1794) [омонимия снята]
8. Григорий Сковорода. Толкование из Плутарха о тишине сердца (1766-1794) [омонимия снята]
9. Н. М. Карамзин. Бедная Лиза (1792) [омонимия снята]
10. Д. И. Фонвизин. Выбор гувернера (1790-1792) [омонимия снята]
11. А. Н. Радищев. Записки путешествия в Сибирь (1790) [омонимия снята]

Text mining toolkit

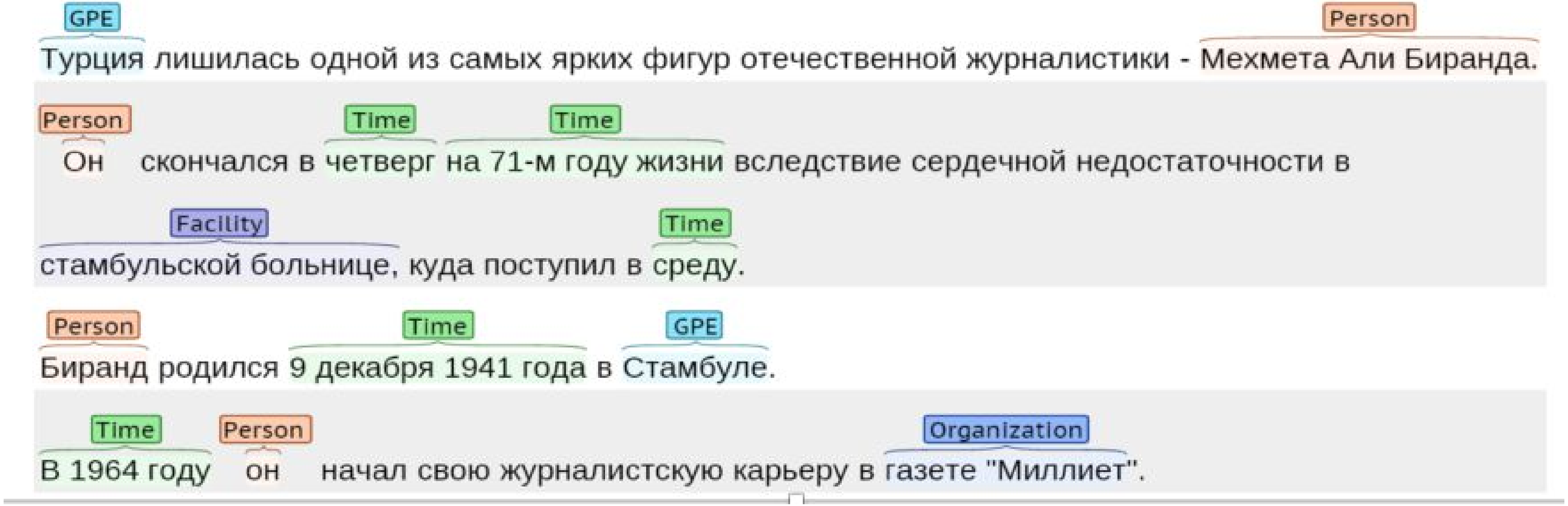
Components for text processing:

- Text Segmenter
- Morphological analyzer
- PoS-tagging
- Lemmatizer
- Chunker
- Syntactic Analyzer
- Named Entity Recognizer
- Event Extractor

Tasks: information extraction (NER, EE)

Open-source toolkit: https://github.com/textocat/textokit-core

Examples

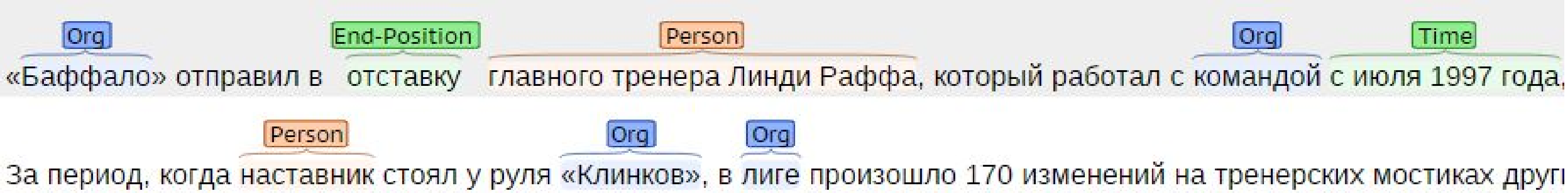


Турция лишилась одной из самых ярких фигур отечественной журналистики - Мехмета Али Биранда.

Он скончался в четверг на 71-м году жизни вследствие сердечной недостаточности в стамбульской больнице, куда поступил в среду.

Биранд родился 9 декабря 1941 года в Стамбуле.

В 1964 году он начал свою журналистскую карьеру в газете "Миллиет".



«Баффало» отправил в отставку главного тренера Линди Раффа, который работал с командой с июля 1997 года,

За период, когда наставник стоял у руля «Клинков», в лиге произошло 170 изменений на тренерских мостиках други

Information Extraction

- Named Entity Recognition (NER)
- Event Extraction (EE)

NER. Original Corpus

Contains:

- 100 documents from different Russian online newsfeeds;
- ~35000 words.

Manually annotated:

- ~1300 Organization mentions;
- ~500 Person mentions.

Paper: *Gareev, Tkachenko, Solovyev, Simanovsky, Ivanov*. **Introducing Baselines for Russian Named Entity Recognition**. CICLING-2013.

NER. Quality

Person F1	**~ 0.79**
Organization F1	**~ 0.55**

Future work:

1) extend the corpus;
2) annotate other NE types: GPE, Locations, Temporal Expressions;
3) implement a Sequential Classification-based Tagger with morphological and chunker features.

Event Extraction

Task: extract event mentions of predefined types and fill their arguments from Russian news texts.

Example event types:

- Merge & Acquisition
- Person Position Change

We follow Automatic Context Extraction

(ACE-2005) definitions mostly.

EE. Experience

Finished the 1-year project in collaboration with HP Russia Labs (Saint-Petersburg)

Elaborated rule-based extraction

- use the NER component;
- match '*event triggers'* in a text;
- apply rules to match arguments around a trigger.

Prepared an evaluation corpus

- 100 docs, ~1500 sentences;
- ~130 event mentions (3 event types) have been annotated so far

EE. Ongoing efforts

Moving from 'linear patterns' matching

towards applying trigger-specific mappings

fill an event argument by a nearby NP that satisfies certain morphological and semantic constraints

The experiments have shown that the latter approach significantly better, **but** requires more sophisticated preprocessing steps:

- PoS-tagging;

- NP recognition.

Current Projects

1. Text mining models and methods for analysis of the needs, preferences and consumer behavior (RSCF grant, 15-11-10019)
2. Computational models and mathematical methods for big data analysis of trends and correlations in society (RFBR grant, 15-29-01173)
3. Mining Hypotheses concerning Drug Discovery and Drug Repurposing (joint work with Cheminformatics Lab of KFU)
4. Pharmacovigilance based on Social Media Posts written in Russian
5. Named Entity Recognition in Legal Documents (joint work with the Faculty of Law of KFU)

Analysis of consumer behavior and user opinions

- Develop, implement and evaluate text mining methods for processing of consumers’ texts:
    - Deep neural networks for prediction of demographic information from medical user reviews;
    - Deep neural networks for bilingual sentiment analysis of short texts;
    - Aspect-based sentiment lexicons constructed with topic modeling;
    - Methods for inferring sentiment-based priors in topic models;
    - Topic models for extracting failures from product reviews;
    - Knowledge-driven Event Extraction in Russian.

Cheminformatics & Text Mining & Data Mining

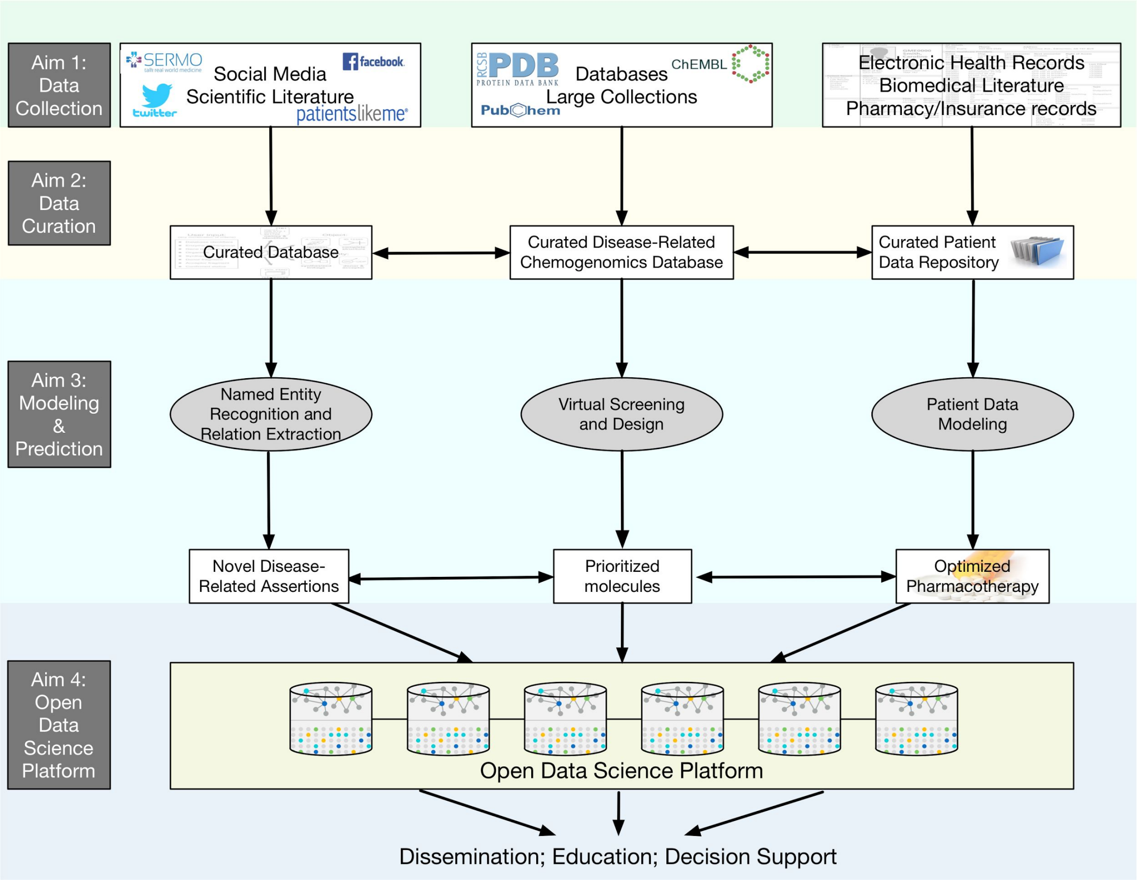
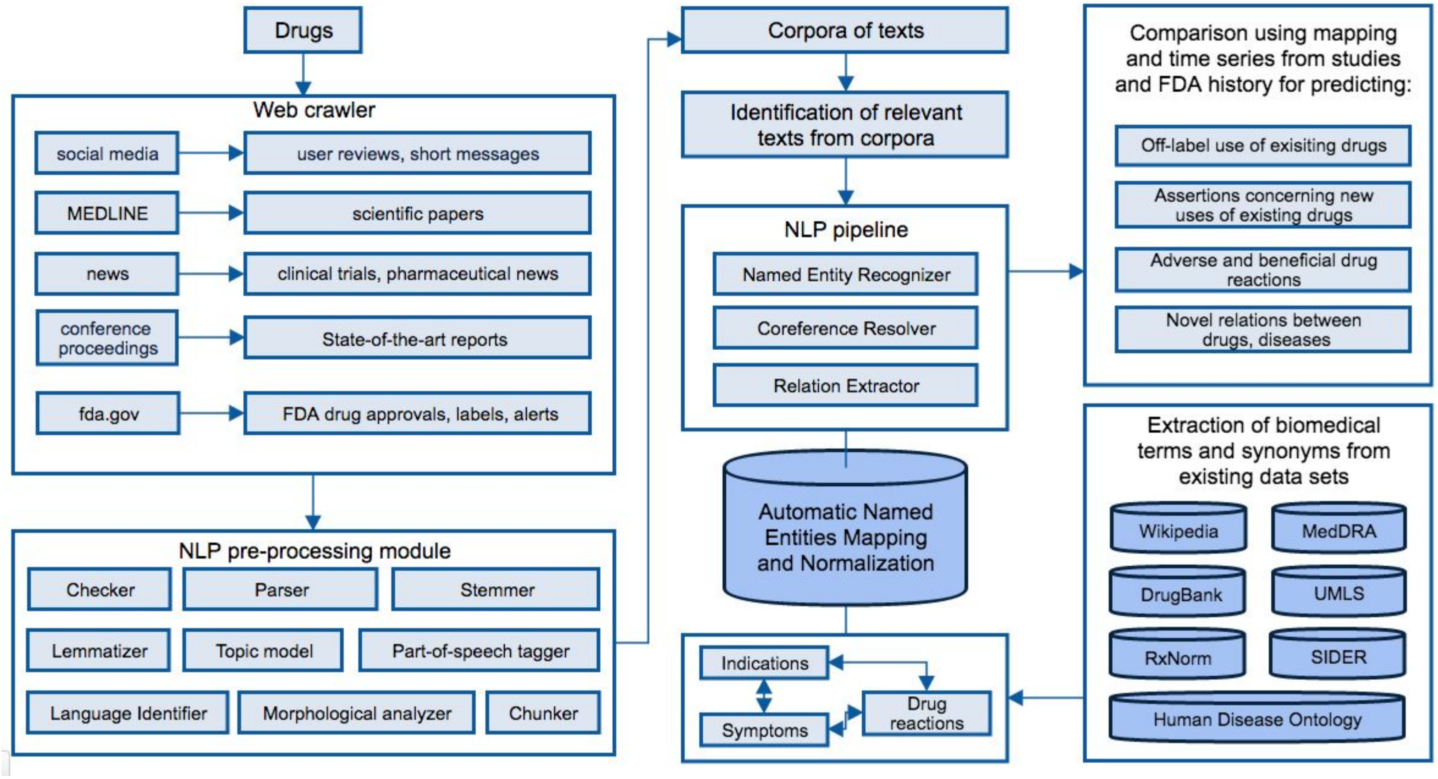

Aim 1: Data Collection
SERMO
facebook
Social Media
Scientific Literature
twitter
patientslikeme
RCSB PDB PROTEIN DATA BANK
Databases
ChEMBL
Large Collections
PubChem
Electronic Health Records
Biomedical Literature
Pharmacy/Insurance records
Aim 2: Data Curation
Curated Database
Curated Disease-Related Chemogenomics Database
Curated Patient Data Repository
Aim 3: Modeling & Prediction
Named Entity Recognition and Relation Extraction
Virtual Screening and Design
Patient Data Modeling
Novel Disease-Related Assertions
Prioritized molecules
Optimized Pharmacotherapy
Aim 4: Open Data Science Platform
Open Data Science Platform
Dissemination; Education; Decision Support

Image credit to
A. Tropsha (UNC USA)

NLP pipeline for Disease Extraction



Drugs

Web crawler

social media → user reviews, short messages

MEDLINE → scientific papers

news → clinical trials, pharmaceutical news

conference proceedings → State-of-the-art reports

fda.gov → FDA drug approvals, labels, alerts

NLP pre-processing module

Checker

Parser

Stemmer

Lemmatizer

Topic model

Part-of-speech tagger

Language Identifier

Morphological analyzer

Chunker

Corpora of texts

Identification of relevant texts from corpora

NLP pipeline

Named Entity Recognizer

Coreference Resolver

Relation Extractor

Automatic Named Entities Mapping and Normalization

Indications

Symptoms

Drug reactions

Comparison using mapping and time series from studies and FDA history for predicting:

Off-label use of exisiting drugs

Assertions concerning new uses of existing drugs

Adverse and beneficial drug reactions

Novel relations between drugs, diseases

Extraction of biomedical terms and synonyms from existing data sets

Wikipedia

MedDRA

DrugBank

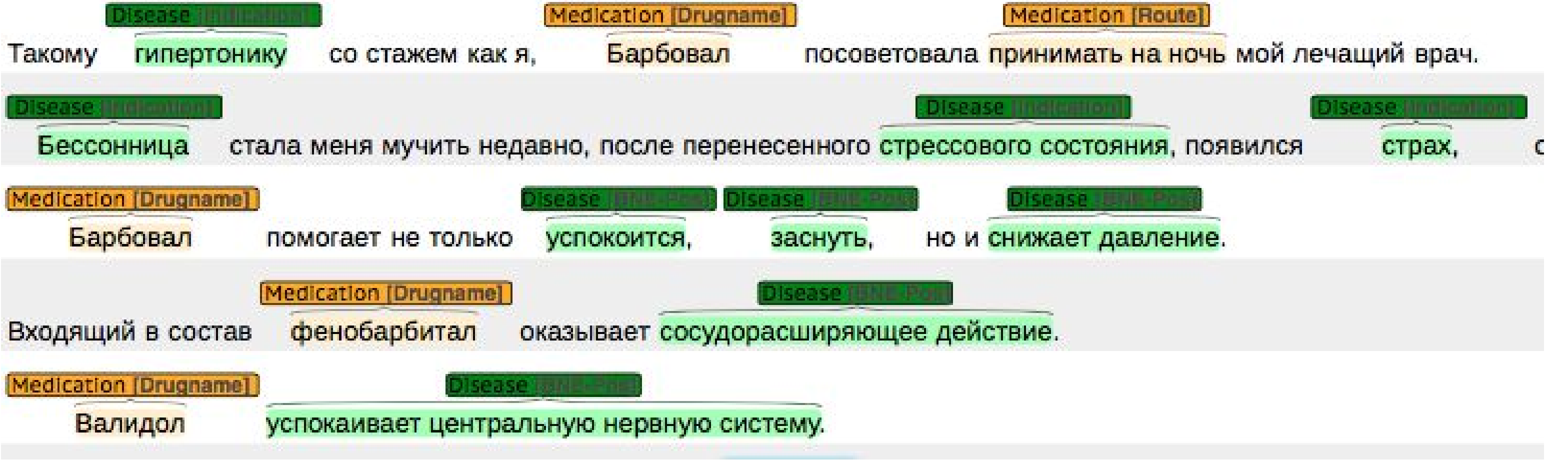
UMLS

RxNorm

SIDER

Human Disease Ontology

Examples



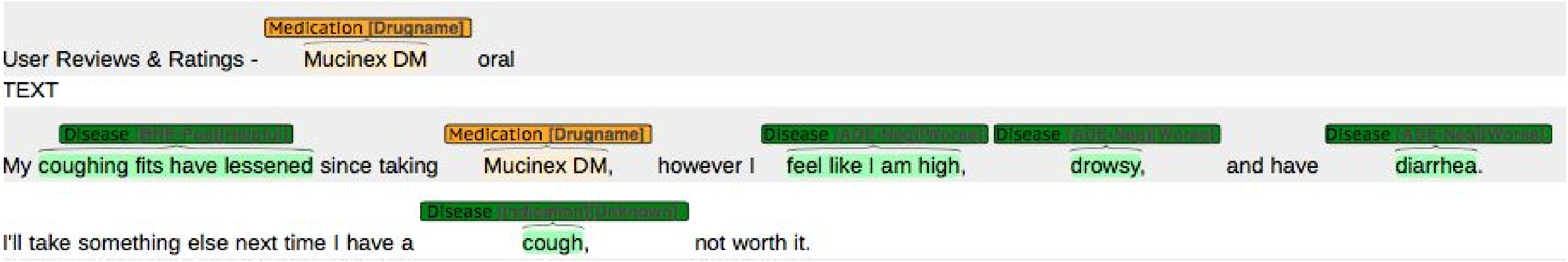
Такому гипертонику со стажем как я, Барбовал посоветовала принимать на ночь мой лечащий врач.

Бессонница стала меня мучить недавно, после перенесенного стрессового состояния, появился страх, с

Барбовал помогает не только успокоится, заснуть, но и снижает давление.

Входящий в состав фенобарбитал оказывает сосудорасширяющее действие.

Валидол успокаивает центральную нервную систему.



User Reviews & Ratings - Mucinex DM oral

TEXT

My coughing fits have lessened since taking Mucinex DM, however I feel like I am high, drowsy, and have diarrhea.

I'll take something else next time I have a cough, not worth it.