Обработка естественного языка и искусственный интеллект

Иван Валентинович Смирнов
ИСА ФИЦ ИУ РАН
Лаборатория «Компьютерная лингвистика и интеллектуальный анализ информации»

Прикладные задачи обработки естественного языка

- Машинный перевод
- Человеко-машинное взаимодействие, диалоговые системы
- Анализ и синтез речи
- Информационный поиск по запросу, вопросно-ответный поиск
- Извлечение информации из текстов
- Поиск похожих текстов и выявление заимствований
- Анализ тональности и эмоциональной заряженности текстов
- Анализ качества текстов
- Резюмирование текстов
- Атрибуция текстов
- И др.

I. Анализ текстов

Лингвистический анализ текстов

- Морфологический анализ
  - определение словарных форм слов
  - установление морфологических признаков
  - снятие морфологической многозначности
- Синтаксический анализ
  - сегментация предложений
  - установление связей между словами
- Семантический анализ
  - определение «смысла» (значения) слов и высказываний
- Дискурсивный анализ
  - установление отношений между высказываниями

Семантика текста

Модели семантики:

- Лексическая семантика – толкование слов
- Дистрибутивная семантика – значения слов через их сочетаемость с другими словами
- Модель СМЫСЛ-ТЕКСТ
- Семантика высказываний
  - Модель семантических фреймов
  - Формальная семантика Монтегю
- И другие

Коммуникативная грамматика (Золотова Г.А.)

- Синтаксема – минимальная семантико-синтаксическая единица русского языка, характеризующаяся морфологической формой, синтаксической функцией и значением
- В конкретном предложении слово выступает в качестве единицы смысла именно как синтаксема. Таким образом, при работе с текстом необходимо оперировать не лексическими единицами, а синтактико-семантическими (синтаксемами)
- Значение синтаксемы передаёт элементарный смысл
- Синтаксический словарь (Золотова Г.А.) описывает синтаксемы с их синтаксическими значениями (ролями) с примерами в контексте

Реляционно-ситуационная модель (Осипов Г.С.)

Примеры значений (семантических ролей):

- Субъект – компонент предикации (*исследование показало перспективность*)
- Объект – подвергающийся воздействию (*сделан выбор направления исследований*)
- Директив – направление движения (*отправиться в Германию*)
- Аблатив – исходная точка движения (*выйти из комнаты*)
- Локатив – компонент со значением местонахождения (*войска сосредоточены в районе Багдада*)
- Каузатив – причина (*гипертония приводит к поражению артерий)*
- Результатив – следствие (*гипертония приводит к поражению артерий*)

Всего выделено 74 роли

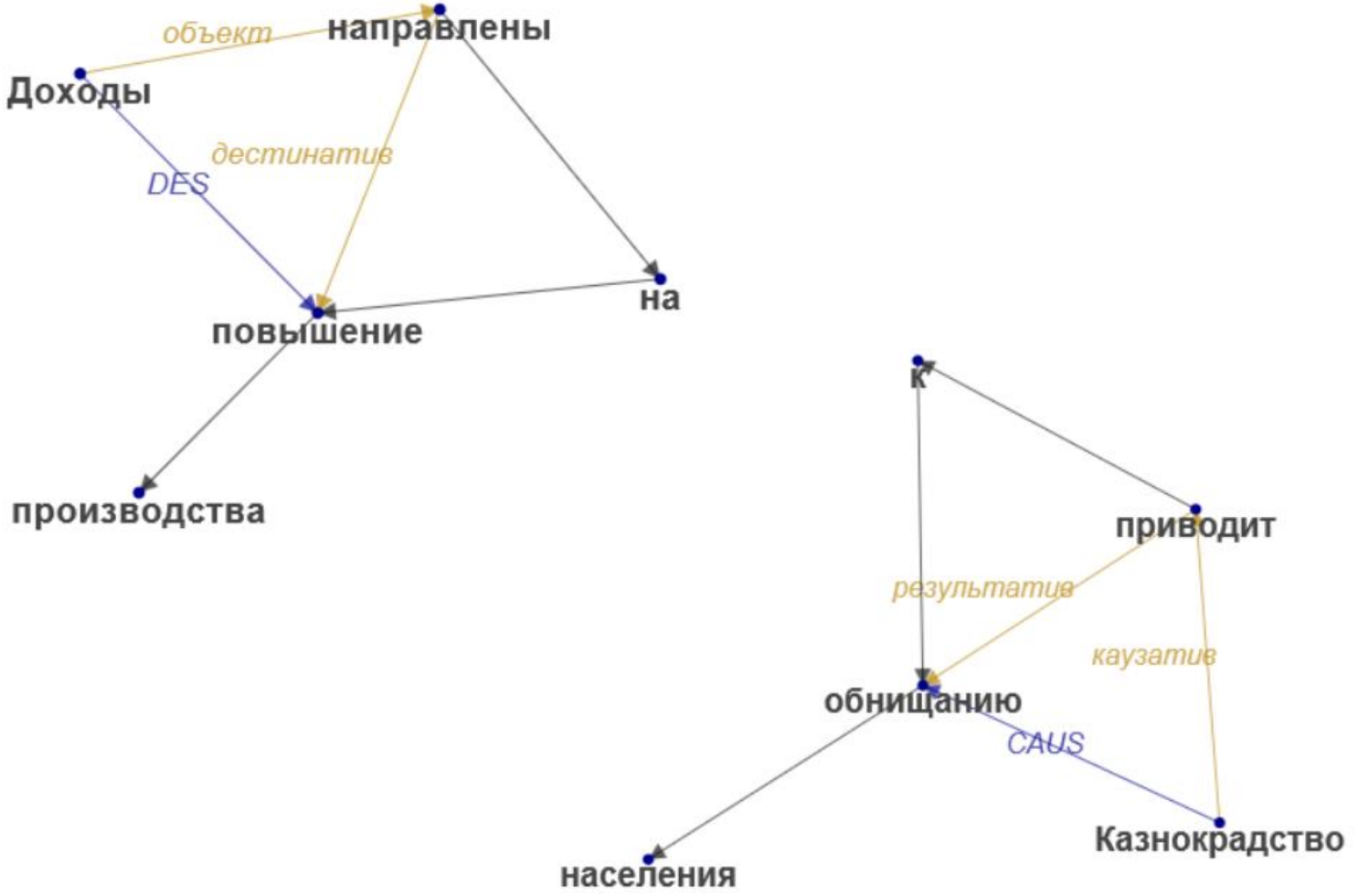
Реляционно-ситуационная модель

Примеры семантических связей

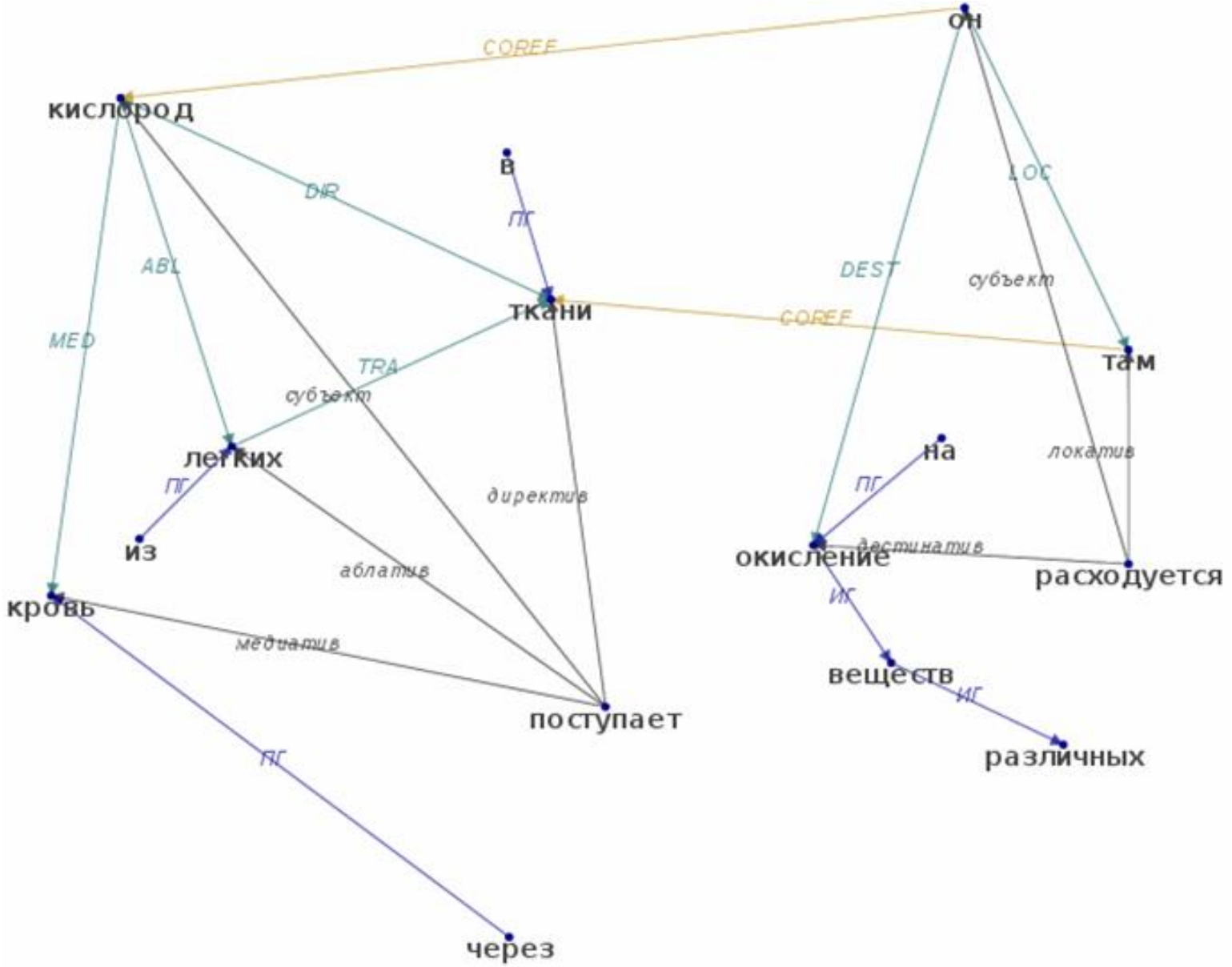
- DIR – директивная связь, в которой один компонент обозначает путь, направление второго компонента (*Владимир Путин отправился* в *США*)
- DEST – дестинативная связь, один компонент которой обозначает назначение для другого компонента (*доходы направлены на повышение производства*)
- LOC – локативная связь, один компонент которой называет местонахождение другого компонента (*В Париже с успехом прошли гастроли Большого театра*)
- CAUS – каузальная связь, один компонент которой обозначает причину проявления другого компонента спустя какое-то время (*Казнокрадство приводит к обнищанию населения*)
- POS – посессивная связь, один компонент которой выражает отношение владения другим компонентом (*Абрамовичу принадлежит ф/клуб «Челси»*)

Всего выделено 30 связей

Представление высказываний



Графовое представление текстов



Задачи поиска и анализ массивов текстов

- Поиск по запросу
- Вопросно-ответный поиск
- Сравнение текстов, классификация и кластеризация текстов
- Поиск заимствований
- Поиск похожих по смыслу текстов

- При решении этих задач необходимо учитывать не просто слова-лексемы, а слова с их семантическими значениями. Предполагается, что это повышает качество решения задач.
- Необходима промышленная реализация анализаторов и систем

Именная синтаксема

- Морфологическая форма – предлог, падеж, категориальный класс (личное, пространственное, признаковое, предметное и т.д.)
- Синтаксическая функция:
  - I - самостоятельное употребление
  - II - употребление в качестве компонента предложения
  - III - присловное употребление в качестве компонента словосочетания
- Синтаксическое значение (роль) – элементарный смысл, передаваемый синтаксемой

- Примеры:
  - <из-за, род.п., пространственное, I, аблатив>: *из-за синих гор*
  - <из-за, род.п., признаковое, I, каузатив>: *из-за пустяка*

Семантический анализ. Словарный подход

Леммы предикатных слов	Отправить, отправлять, направить, направлять, послать, посылать, сослать, выслать, слать (класс: акциональные).	
Семантические роли (синтаксические значения)		
Субъект (инициатор действия)	**КСК**	**Морфологические признаки**
	Личное	**Предлог** / **Падеж**
		- / Именительный
Объект (компонент, подвергающийся действию)	**КСК**	**Морфологические признаки**
	Любой	**Предлог** / **Падеж**
		- / Винительный
Директив (направление движения, ориентированного действия или положения предмета)	**КСК**	**Морфологические признаки**
	Локатив Предметное	**Предлог** / **Падеж**
		В / Винительный
		За / Винительный
		На / Винительный

Семантические связи

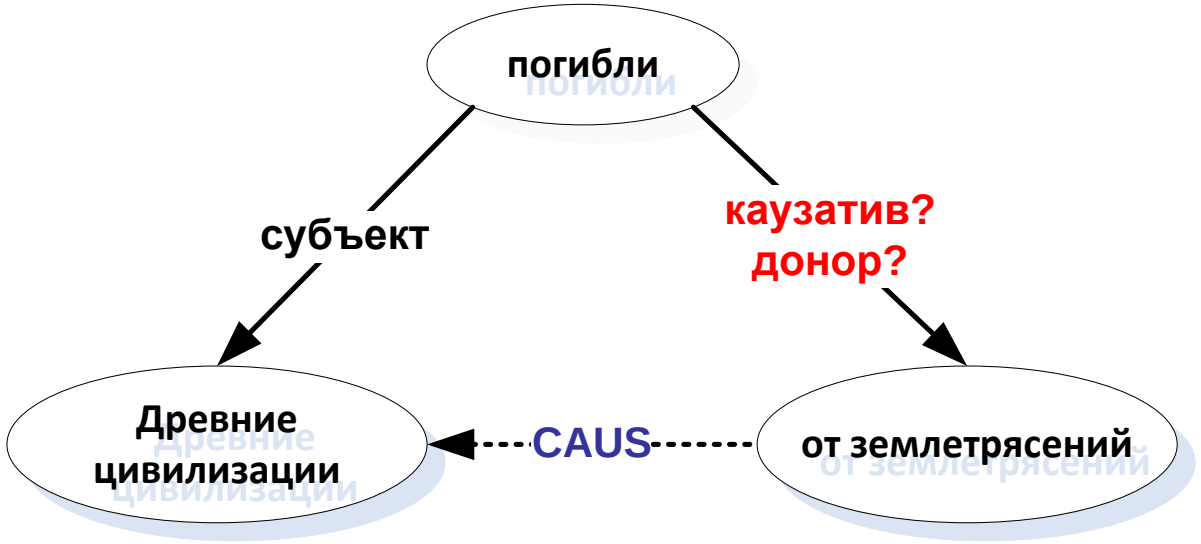
Семантическая связь	**Роль 1**	**Роль 2**
OBJ	Субъект	Объект
DIR	Объект	Директив

- ≈ 5 лет работы лингвистов
- ≈ 2,6 тыс. статей
- ≈ 3,5 тыс. предикатных слов (глаголов и отглагольных существительных)

Метод установления значений по словарю

- Каждой предикатной конструкции сопоставляется множество словарных статей (ролевых структур) из сем. словаря
- При сопоставлении признаков роли в словарной статье с признаками аргумента определяется вес роли $\in [0,1]$
- Особенности сопоставления:
  - Аргументы могут иметь несколько признаков (например, несколько КСК или падежей из-за неснятой омонимии)
  - Одни совпадения признаков весомее других
  - Решается задача оптимизации распределения ролей с весами в предложении
- F-мера $\approx$ 73%

А если глагола нет в словаре? Или в предложении нет глагола?



Пример полисемии синтаксемы:

<от, родительный, предметные> может иметь значения:

- темпоратив – начало отсчёта во времени
- каузатив – причина воздействия
- сурсив – источник информации или восприятия
- деструктив – объект разрушающего воздействия
- абстинатив – нежелательное действие субъекта …

Правила лингвиста

Если встречаем последовательность

кто/ что – из чего, при этом

1) в позиции компонента *кто/ что* находим *личное*, реже – *предметное* существительное;

2) в позиции компонента *из чего* находим *локативное* или *предметное* сущ.,

то

1) компонент *кто/ что* следует считать субъектом;

2) компонент *из чего* следует считать предикатом со значением *аблатива*.

Можно построить такие правила автоматически. Нужны размеченные данные.

Автоматический вывод правил из словаря синтаксем

Принцип индуктивного ДСМ-рассуждения:

> *Если какое-то обстоятельство постоянно предшествует наступлению исследуемого явления, в то время как иные обстоятельства изменяются, то это обстоятельство есть, вероятно, причина данного явления*

Расширения ДСМ-метода (Финн В.К.):

- Введено понятие синтаксемы в контексте
- Введена операция вычисления сходства для синтаксем в контексте
- Показано, при каких условиях введенные понятия могут использоваться для порождения правил

Сравнение с правилами лингвиста

- Точность = 0.83
- Полнота = 0.58

Примеры правил

Если

встречается синтаксема в падеже <родительный> с предлогом <для>, имеющая категориальный класс <личное>, а до неё встречается синтаксема в падеже <именительный>, имеющая категориальный класс <предметное>

То

первая синтаксема имеет значение <дестинатив – назначение предмета>

Фрагмент обучающей выборки:

Пример 1:

LABEL=дестинатив

ЦЕЛЕВАЯ СИНТАКСЕМА: для тебя; КСК: личное

СОСЕДНЯЯ СИНТАКСЕМА: Все; ПРЕДЛОГ: ; ПАДЕЖ: именительный, винительный; КСК: предметное; ПОЗИЦИЯ: до

===КОНТЕКСТ: и песни , и силы - Все для тебя

Пример 2:

LABEL=дестинатив

ЦЕЛЕВАЯ СИНТАКСЕМА: для различных рачков; КСК: личное

СОСЕДНЯЯ СИНТАКСЕМА : пища; ПРЕДЛОГ: ; ПАДЕЖ: именительный; КСК: предметное; ПОЗИЦИЯ: до

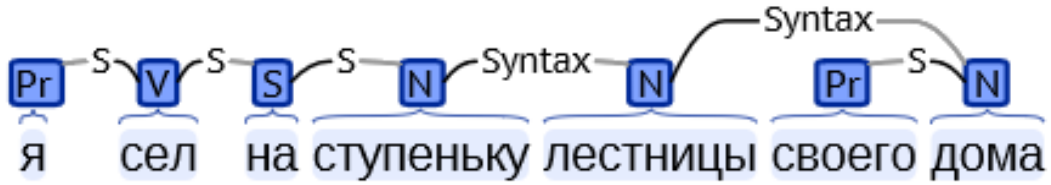
===КОНТЕКСТ: Эти растения - пища для различных рачков

Проблемы синтаксиса

- Ошибки синтаксиса приводят к ошибкам семантики
- Нужен глубокий синтаксис, чтобы обрабатывать сложные предложения
- Пример синтаксического разбора АОТ

Обучаемый синтаксис

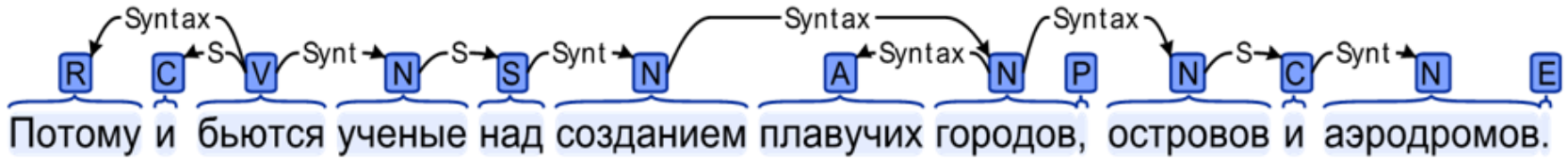
- Синтаксически размеченный корпус русского языка «СинТагРус» (версия 2012 года):
  - 53 439 предложений
  - 774 373 токенов без учета пунктуации
  - проблема перевода в структурированный вид (БД)
  - знаки пунктуации не считаются токенами
- MaltParser

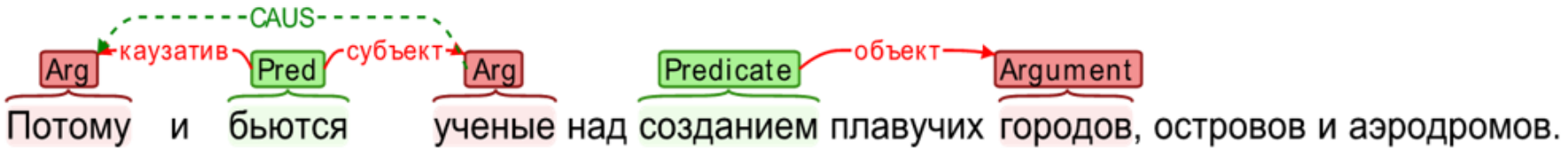


- F-Мера ≈ 0.89

Синтактико-семантический анализ. Идея

- Синтаксис и семантика тесно взаимосвязаны
- Для построения правильной синтаксической структуры предложения необходима семантика





- Семантический и синтаксический анализ можно выполнять одновременно, на одних и тех же структурах данных
- Ролевая связь между предикатным словом и аргументом является разновидностью синтаксической связи с пометкой роли

Проверка на корпусе

- Семантически размеченный подкорпус «СинТагРус» (ИСА РАН, год работы 2-х лингвистов):
  - более 1 700 предложений
  - около 29 000 токенов без учета пунктуации
  - около 3 000 предикатных конструкций
  - около 4 000 аргументов, с установленными ролями
- Собственный разметчик

Экспериментальное исследование системы семантико-синтаксического анализа на задаче определения ролевых структур высказываний

- Сравнивались две системы:
  - **MaltParser + Сем**антический анализатор. Синтаксический и семантический анализ выполняются раздельно
  - **Сем.-син.** – система семантико-синтаксического анализа, метод проверки исправлений на основе машинного обучения
- Качество определения ролевых структур высказываний:

Система	p,%	r,%	F_1,%
MaltParser + Сем	89,6	61,0	72,6
Сем.-син.	**89,6**	**62,7**	**73,8**

- Устанавливает **дополнительно более 10 %** синтаксических связей для наиболее сложных случаев, которые не были обнаружены MaltParser
- Точность исправлений синтаксических связей **≈ 85%**

Проблемы

- Составление правил или словарей трудоёмко
- Для обучения нужно много размеченных текстов
- Разметка корпусов очень трудоёмка, качество разметки может быть низким
- Составленные вручную правила, словари или обученные на корпусе алгоритмы могут плохо работать на текстах других жанров или предметных областей. Для новой предметной области необходимо пополнять словарь.
- Алгоритмы, обученные на корпусах с эталонной разметкой, плохо работают на реальных текстах (в том числе из-за ошибок анализаторов)
- Пример работы словарного анализатора на эталонных и реальных данных:

Признаки	Recall, %	Precision, %	F_1, %
Морфология и синтаксис из СинТагРус	82.5	94.5	88.1
Морфология СинТагРус + реальный синтаксис	67.3	88.7	76.5

Обучение с частичным привлечением учителя (semi-supervised machine learning)

Принципы

- Используем небольшое количество размеченных данных
- Применяем самообучение (self-learning), когда классификатор обучается на результатах своей работы
- Применяем совместное обучение (co-training), когда два классификатора, отличающихся парадигмами классификации / признаками / источниками данных, поочередно обучаются на результатах работы друг друга
- Используем кластеризацию с частичным обучением – когда в кластеризуемом множестве присутствуют объекты, класс которых известен

Алгоритм самообучения для установления семантических ролей

1. Применяем словарный семантический анализатор для разметки СинТагРус семантическими ролями на основе «эталонных» морфологии и синтаксиса
2. Выполняем морфологическую и синтаксическую разметку СинТагРус своими «реальными» анализаторами
3. Обучаем семантический анализатор на полученных «реальных» морфологии и синтаксисе и словарной семантике
4. Размечаем обученным семантическим анализатором СинТагРус, сравниваем полученную разметку с разметкой на шаге 1.
5. Убираем примеры (аргументы и предложения), для которых результаты не шаге 4 и 5 сильно отличаются
6. Повторяем с шага 3 на очищенной разметке до тех пор, пока есть прирост качества на заранее отложенной контрольной части СинТагРус с семантической разметкой

Предварительные результаты установления ролей при самообучении

Dictionary-based parser

Test set	Recall, %	Precision, %	F1, %
All features are DIRTY	65.9	85.7	74.5

Data-driven parser

Test set	Recall, %	Precision, %	F1, %
All features are DIRTY	66.5	85.4	74.8

Data-driven parser with training set enhancement

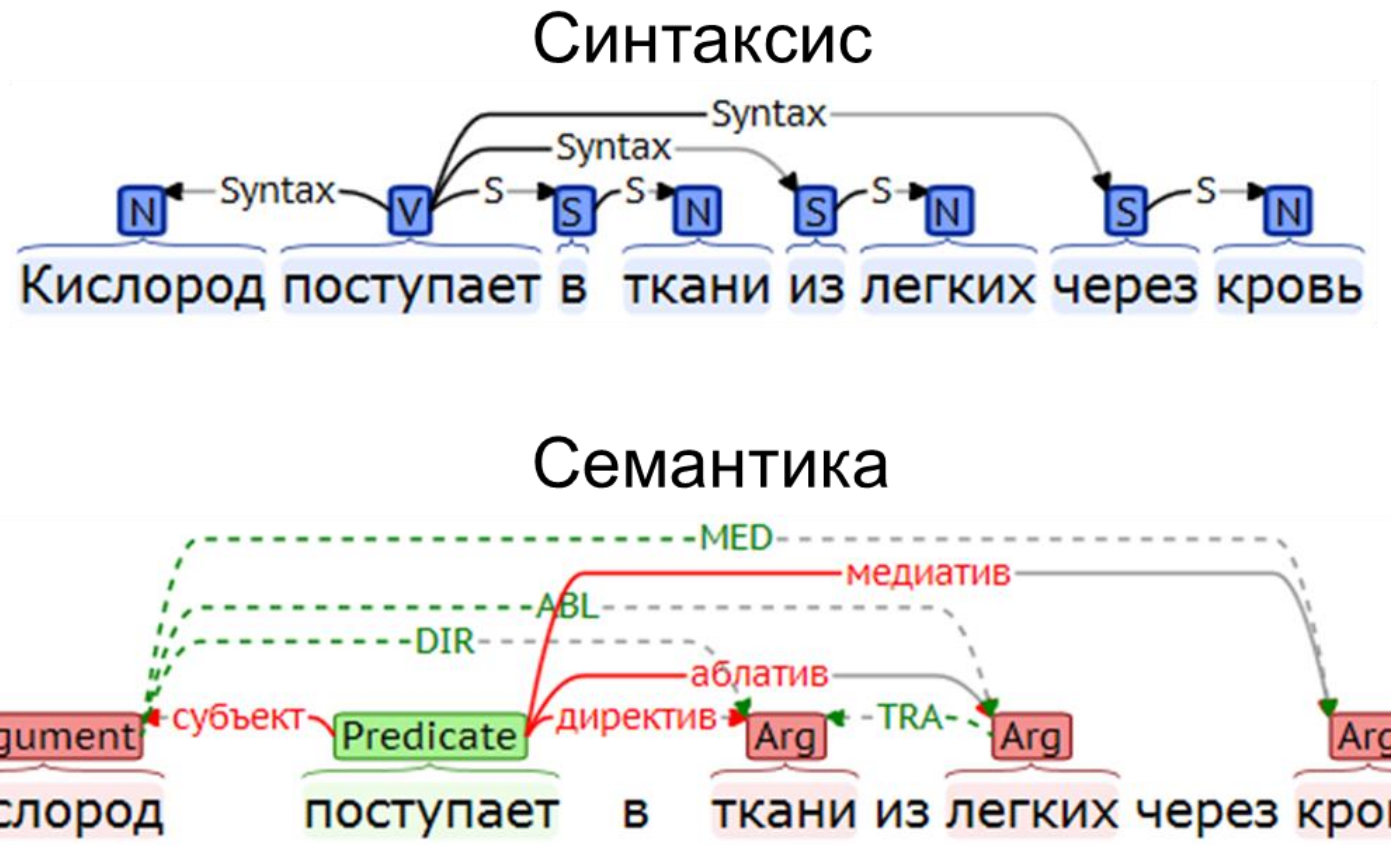
Test set	Recall, %	Precision, %	F1, %
All features are DIRTY	66.7	85.9	75.1

Обучение без учителя (Unsupervised Learning)

- Без ручного построения правил, без использования обучающих корпусов
- Кластеризация конструкций любой сложности на основе их лингвистических признаков и контекста
- Кластеризация пар конструкций для определения отношений на основе их лингвистических признаков и контекста
- Использование векторных представлений
- Функционал качества задаётся таким образом, чтобы порождаемые кластера приближенно соответствовали заданным типам конструкций (сущность, аргумент и т.д.) и отношений (семантических, риторических и др.)
- Извлекаемые группы конструкций и отношений **не** именуются автоматически
- Приложение: ускоренное создание крупных размеченных корпусов. Экспертам достаточно лишь провести именование сравнительно небольшого набора полученных групп конструкций и отношений

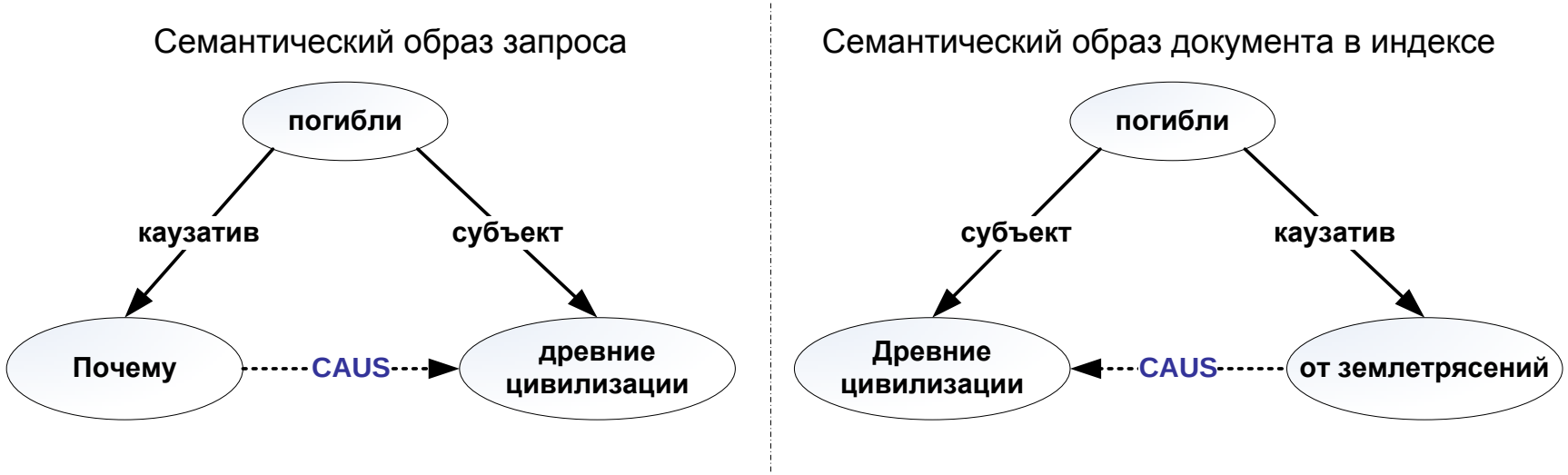
Примеры результатов анализа

http://nlp.isa.ru



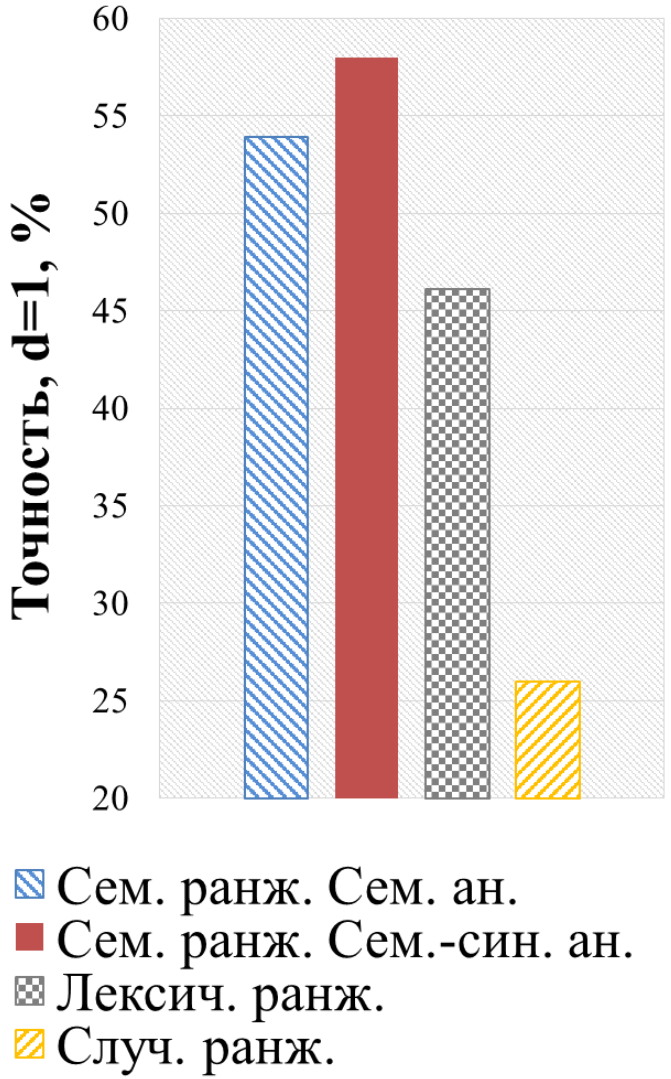
Такие структуры укладываются в поисковые индексы

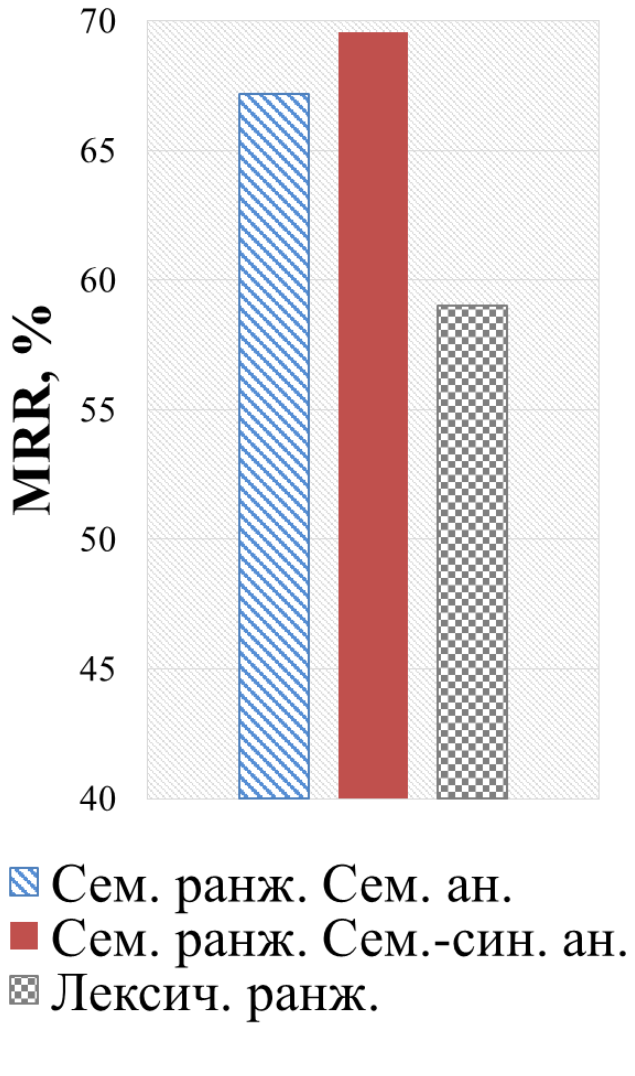
Вопросно-ответный поиск



Алгоритм показал лучшие результаты по всем параметрам качества вопросно-ответного поиска на РОМИП-2010

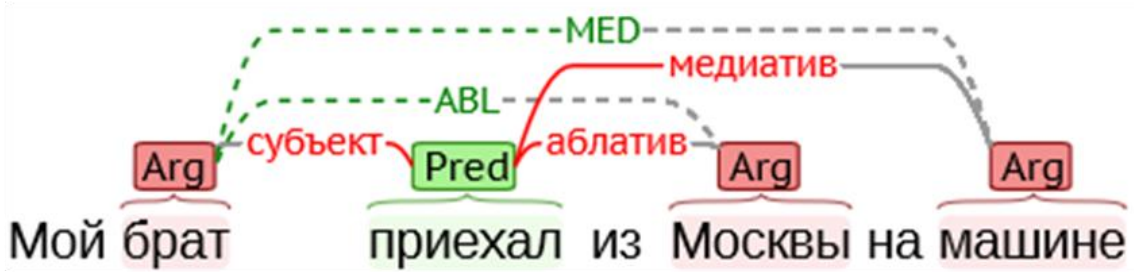
Оценка вклада семантики





Семантическое сходство текстов

Семантическая структура текста источника



Семантическая структура заимствования

При перефразировании семантика текста сохранилась

Системы

- Exactus – exactus.ru – семантический поиск - демо
- Exactus Expert – expert.exactus.ru – интеллектуальный поиск и анализ научной информации
- Exactus Like – like.exactus.ru – поиск заимствований в научных текстах
- Exactus Patent – patent.exactus.ru – патентный поиск и анализ
- TextAppliance – textapp.ru – программно-аппаратный комплекс интеллектуального поиска и анализа больших массивов текстов

Отличительные особенности:

- Интеграция статистических и лингвистических подходов для обработки больших массивов текстовой информации

Анализ дискурса

Дискурс – уровень языка, выходящий за границы предложения и представляющий текст в виде древовидной иерархической структуры.

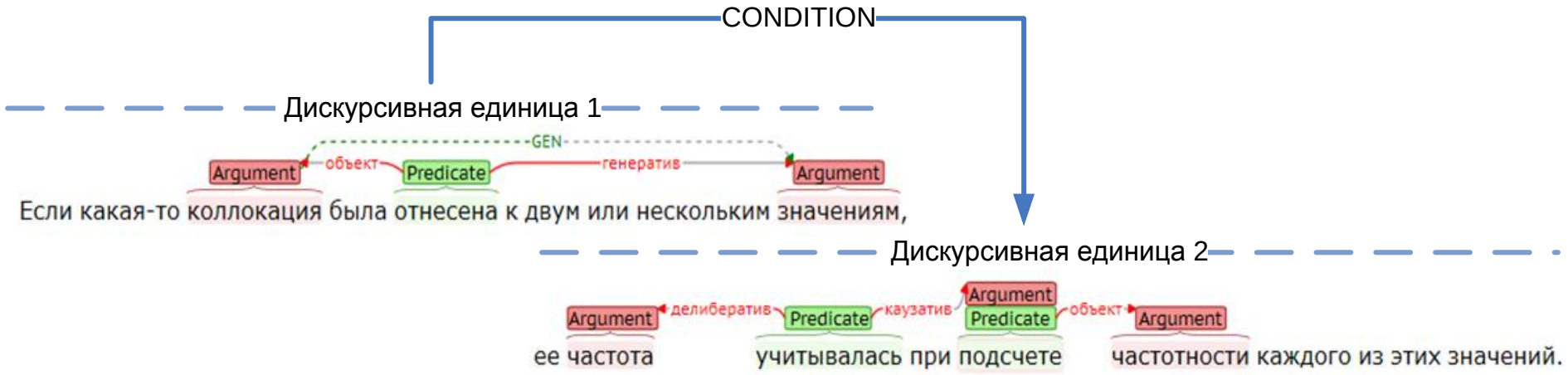
Основные понятия: дискурсивные единицы (клаузы и предложения) и семантические отношения между ними

Применяется при решении многих задач, таких как: автоматическое реферирование текстов, анализ тональности, жанровая классификация, вопросно-ответный поиск и др.

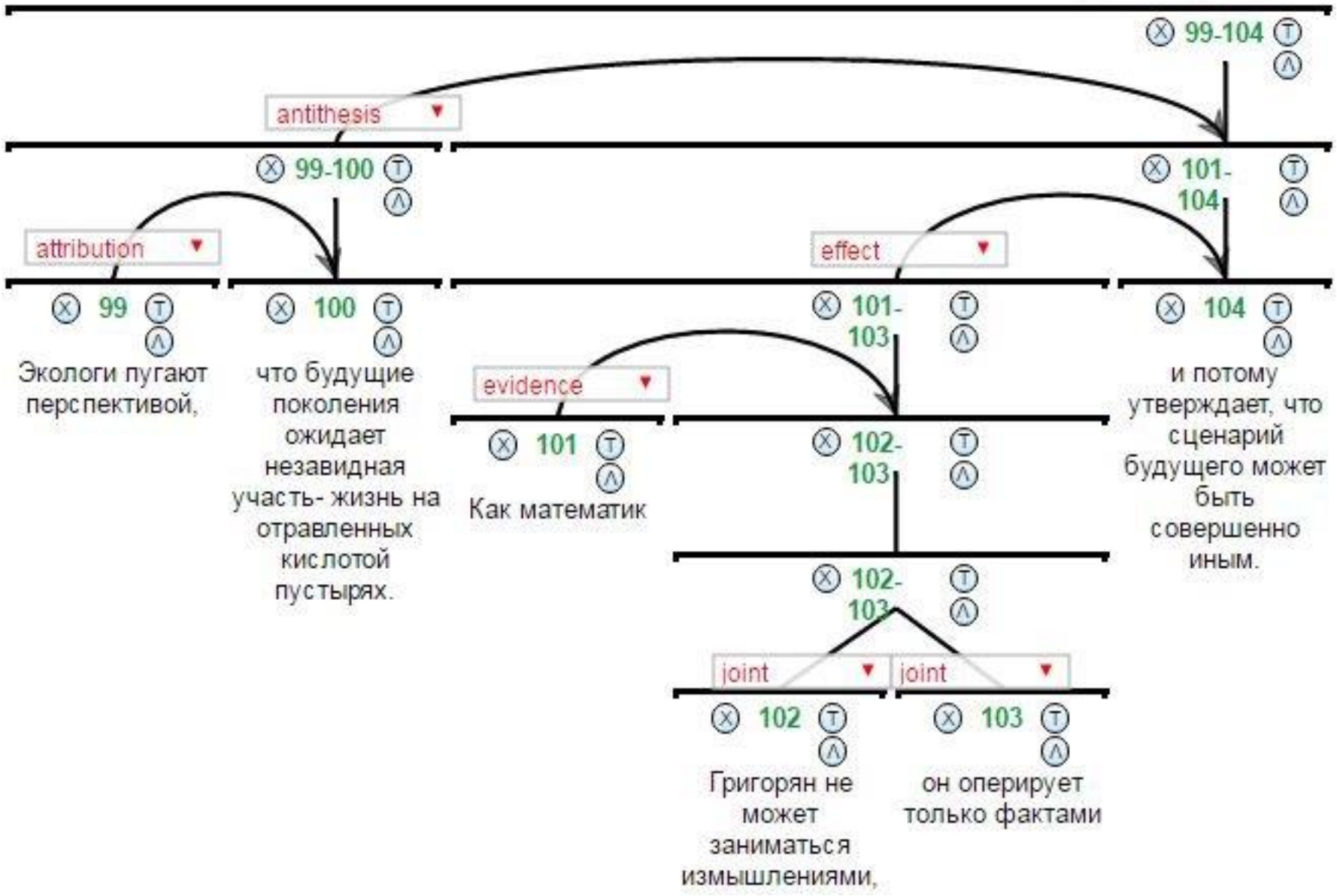
Учет дискурса повышает качество решения задач обработки языка

Теория риторических структур (Mann, 1988)

- Каждая единица дискурса связана дискурсивным отношением по крайне мере с одной другой единицей.
- Список отношений включает 23 отношения. Например: причина, следствие, антитеза, детализация, оценка, обстоятельство, контраст, сравнение, переформулировка и др.



Пример дискурсивной структуры



II. Извлечение информации из текстов

Извлечение информации из текстов

- Трансформация неструктурированной информации в структурированную
- Выявление в текстах различных типов сущностей и отношений между ними для различных предметных областей
- Сохранение извлеченной информации в реляционной базе данных для последующего анализа
- Использование семантической информации

Анализ научных текстов

- Даны научные тексты на русском языке
- Необходимо
  - Извлечь из текстов дефиниции и формулировки научных результатов
  - Выделить в тексте зоны аргументации, соответствующие постановке проблемы, обзору других работ, описанию предложенных в работе методов, результатов экспериментов и т.п.
  - Оценить содержание в тексте научной и псевдонаучной лексики
  - Оценить грамотность текста
  - Выставить общую оценку качества научного текста на основе указанных оценок

Примеры шаблонов для выделения дефиниций из научных текстов на русском языке

Правило	Примеры
ЧР(Сущ.) && Сем.роль(**эстиматив**) + Л(«называться»)	***Перигелием*** называется точка орбиты небесного тела, где оно максимально сближается с Солнцем.
Сем.роль(**делиберативе**) + ПС(«определять») + Л(«как»)	***Аксиоматический метод*** традиционно определяется как такой способ дедуктивного построения научной теории, когда ее основу составляют лишь некоторые, принятые без доказательств положения – аксиомы.

Обозначения: «Л» – лемма; «Сем.роль» – семантическая роль; «ЧР» – часть речи; «ПС» – идентификатор предикатного слова.

Оценка метода извлечения определений

- Корпус с разметкой (с использованием BRAT, 6 месяцев):
  - >72 000 токенов
  - >300 определений

Анализатор	***p*,%**	***r*,%**	***F_1*,%**
Сем	82,6	68,3	74,8
Сем.-син.	**82,7**	**68,7**	**75,0**
Без сем.	81,6	58,3	68,0

- Правила, использующие семантические роли, обрабатывают существенное количество случаев определения терминов в научных публикациях
- Семантические роли упрощают построение правил для извлечения определений и авторских терминов

Оценка качества научного текста

Общая оценка документа: научный

Оценка соответствия текста документа формальным требованиям: полностью соответствует

Доля общенаучной лексики:	49%.
Доля ненаучной лексики:	1%.
Список цитируемой литературы:	Список литературы присутствует.
Описание задачи исследования:	присутствует
Описание методов исследования:	присутствует
Выводы исследования:	присутствуют

Количество речевых дефектов: незначительное

Количество нарушений падежного согласования:	0
Количество нарушений синтаксической связности:	низкое
Количество нарушений согласования однородных существительных и управляющего слова:	2
Содержание плеоназмов:	низкое

ФОРМУЛИРОВКИ РЕЗУЛЬТАТОВ:

В данной статье предложен общий методологический подход к созданию моделей поведения информационных систем в условиях катастрофических воздействий, а также оговорены основные принципы и допустимые ограничения при построении таких моделей.

ФОРМУЛИРОВКИ ОПРЕДЕЛЕНИЙ:

1. Под *катастрофами* подразумеваются не только пожар, наводнение или землетрясение, но и возможные непредвиденные сбои в работе служб, разрушение данных или повреждение ЦОД, например, в случае разрыва телекоммуникационных линий, возникшего в результате проведения ремонтных работ, умышленной диверсии или саботажа или просто по халатности и ошибочных действий обслуживающего персонала.
2. Процесс создания такой модели будем называть *моделированием катастрофоустойчивости информационной системы*.

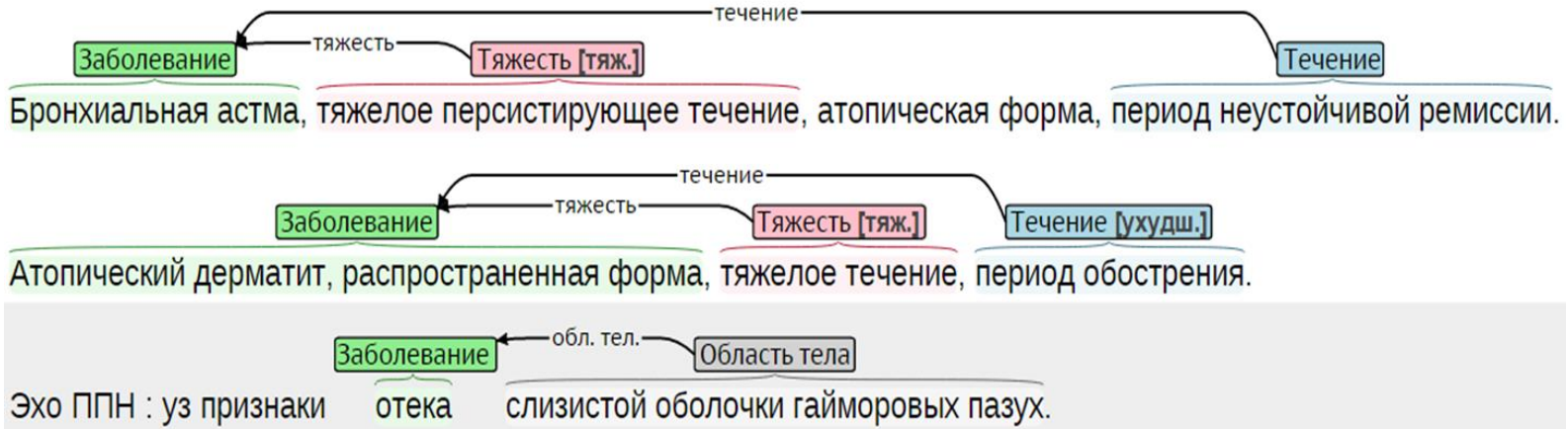
Анализ медицинской информации

Задача

- Даны клинические записи, содержащие:
  - числовую информацию – возраст, пол, результаты анализов и т.п.
  - текстовую информацию – анамнезы, осмотры, эпикризы и т.п.
- Необходимо разработать систему анализа клинической информации для решения следующих задач:
  - Автоматическая диагностика хронических заболеваний у детей
  - Выявление наиболее значимых для диагностики признаков заболеваний
  - Выявление скрытых зависимостей в клинических данных
- Клинические тексты содержат много полезной информации, которую необходимо учитывать для эффективного решения задач

Особенности

- Что выделяем из текстов
  - упоминания заболеваний, симптомов, лекарств, медицинских процедур и др.
  - отсутствие заболевания у пациента
  - заболевание относится не к пациенту (а, например, к его родственнику)
  - тяжесть протекания заболевания
  - сопоставление заболеваний и областей тела, к которым относятся заболевания
- Ручная разметка (BRAT)



- Используем медицинские тезаурус UMLS Metathesaurus, ГРЛС и др.

Корпус

- 120 деперсонализованных историй болезни пациентов с аллергическими, ревматическими и нефрологическими заболеваниями, а также болезнями органов дыхания. Включают текстовые разделы на русском языке:
  - эпикризы
  - рекомендации и отчеты, фиксирующие результаты различных медицинских обследований: УЗИ, ЭКГ, рентгеновские исследования
  - и др.
- Размечено более 18 000 сущностей, более 12 000 атрибутов и связей (6 месяцев, 2 медика)
- По каждому типу извлекаемых данных проводились отдельные эксперименты с применением различных методов обучения

Результаты

Таблица 1. Результаты экспериментальной оценки методов извлечения заболеваний

Метод	Полнота, %	Точность, %	F_1-мера, %
Разработанный метод	72,8	95,1	82,4
Базовый 1	84,9	9,3	16,7
Базовый 2	69,8	99,2	81,9

Таблица 2. Результаты экспериментальной оценки методов распознавания конструкций, указывающих на отсутствие заболевания и на то, что заболевание не относится к пациенту

Метод	Полнота, %	Точность, %	F_1-мера, %
Определение атрибута «отрицание»	98,7	95,3	97,0
Определение атрибута «не пациент»	90,9	96,8	93,8

Таблица 3. Результаты экспериментальной оценки методов извлечения атрибутов «тяжесть заболевания» и «течение заболевания», а также метода установления связей между областями тела и заболеваниями

Задача	Классификатор	Полнота, %	Точность, %	F_1-мера, %
Извлечение атрибута «тяжесть заболевания»	Случайный лес	93,6	82,6	87,5
Извлечение атрибута «течение заболевания»	Линейный SVM	92,3	99,2	95,7
Установление связей между областями тела и заболеваниями	RBF SVM	91,4	76,6	83,3

Shelmanov A. O., Smirnov I. V., Vishneva E. A. Information extraction from clinical texts in Russian // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2015). Issue 14 (21). – 2015. – V1. – pp. 537-549.

Диагностика заболеваний

Нозология	Метод	F_1	
		Только числовые признаки	Текстовые и числовые признаки
Бронхиальная астма	Деревья решений	0,61	**0,86**
	Случайный лес	0,73	**0,98**
	Градиентный бустинг на деревьях решений	0,65	**0,81**
Юношеский артрит	Деревья решений	0,76	**0,90**
	Случайный лес	0,89	**0,95**
	Градиентный бустинг на деревьях решений	0,84	**0,97**

А.А. Баранов, Л.С. Намазова-Баранова, И.В. Смирнов, Д.А. Девяткин, А.О. Шелманов, Е.А. Вишнева, Е.В. Антонова, В.И. Смирнов. Технологии комплексного интеллектуального анализа клинических данных // Вестник РАМН, 2016, №2. – С.160-171.

Анализ медицинской информации

Задача

- Даны научные медицинские статьи на английском языке, описывающие клинические испытания
- Необходимо разработать систему составления мета-анализов, обобщающих результаты нескольких клинических испытаний с выявлением причинно-следственных закономерностей
- Конкретнее: что является причиной успешности лечения пациентов различными типами дендритноклеточных вакцин?

Проблема

- Выделить из текстов детали клинических испытаний

Мета-анализы

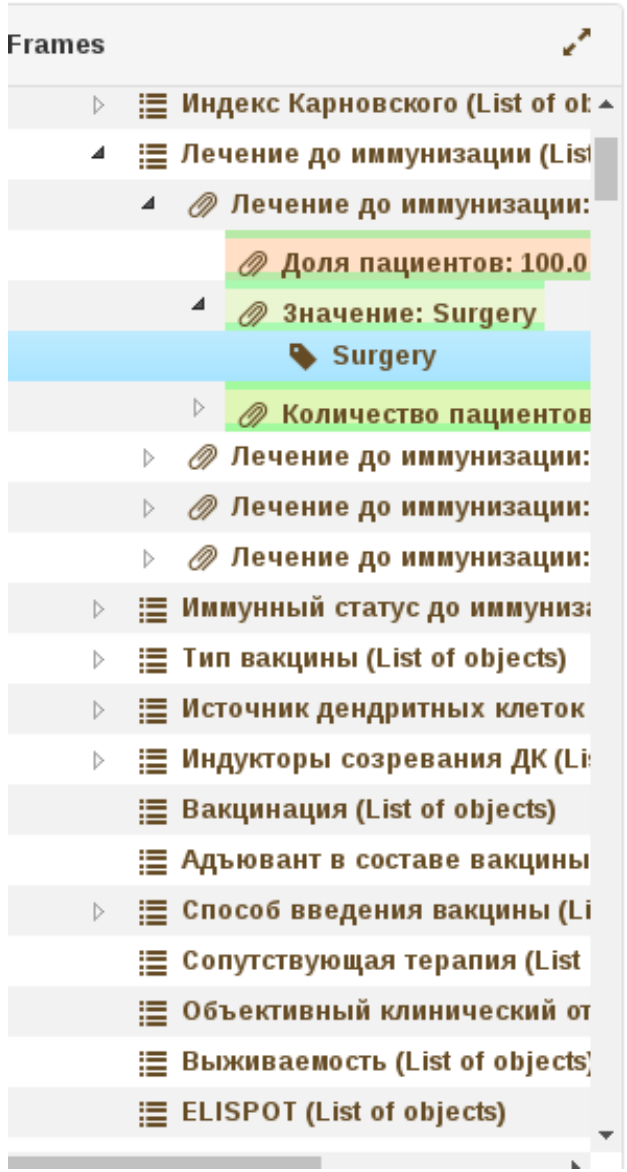
- Что выделяем из текстов

I) Данные о выборке: – Количество пациентов	II) Данные пациента: – Возраст – Пол – Раса – Гаплотип – Иммунный статус до иммунизации
III) Заболевание: – Диагноз – Стадия заболевания – Индекс ECOG до лечения – Индекс Карновского до лечения – Лечение до иммунизации	IV) Схема лечения: – Тип вакцины – Источник дендритных клеток – Индукторы созревания – Вакцинация – и др.
V) Результаты лечения: – Объективный клинический ответ – Выживаемость – Результат – Срок дожития/наблюдения	– Единицы измерения – ELISPOT – Антиген – и др.

- Проблемы: извлекаемые из разных частей текста данные необходимо привязать к одной сущности; между сущностями могут быть отношения

Собственный разметчик

- Позволяет описать сложные составные объекты для извлечения
- Позволяет описать сложные атрибуты объектов
- Позволяет задать связи типа агрегация и композиция между объектами
- Учитывает дальние связи между извлекаемыми конструкциями
- Сохраняет исходное форматирование файлов с расширениями .doc и .pdf
- Интегрирован с cTakes
- Анализ таблиц



Content

a dilution of 1/100 and incubated for 1 hr at room temperature. After unbound complex was removed, peroxidase staining was performed using the substrate 3-amino-9-ethylcarbazole (Sigma). Spots appeared within 4 to 5 min. The color reaction was stopped, and numbers and areas of resulting spots were determined with the use of computer-assisted video image analysis (Herr *et al.*, 1997).

Melanoma-inhibiting activity (MIA) assay

Bosserhoff *et al.* (1997) have demonstrated that MIA can be a useful marker of tumor progression during follow-up of melanoma patients and in monitoring therapy of advanced disease. Human MIA was measured by a 1-step ELISA (Roche, Mannheim, Germany), following the instructions of the manufacturer. The assay is sensitive up to 0.1 ng MIA/ml.

Female	5
Prior therapy	
Surgery	14
Chemotherapy[1]	3
Chemo-/immunotherapy[2]	9
Immunotherapy[3]	5
Metastatic site	
Skin	2
Lymph node	6
Soft tissue	2
Lung	7
Liver	5
Bone	1

[1] Dacarbazine.–[2]Dacarbazine + vinblastine + cisplatin + IL-2 IFN-α2a.–[3]IFN-α2a.

TABLE II – CLINICAL RESULTS

Patient number	HLA-A	Metastatic site	Number of vaccinations	Number of DCs injected	Clinical course (months)
1	HLA-A1	LN, skin	8	5×10^6, 1×10^7	NC[1] (3)
2	HLA-A2	LN	12	5×10^6, 1×10^7	NED[2] (19)

Результаты классификации пациентов

- Корпус:
  - 71 статья
  - > 70 атрибутов
  - выделено 927 групп пациентов

Результаты теста разделимости пациентов с заданными объективными клиническими ответами

Объективный клинический ответ	Число положи-тельных примеров	F1-мера		Точность		Полнота	
		Раздели-мость	3-КВ	Раздели-мость	3-КВ	Раздели-мость	3-КВ
Стабильная болезнь	314	0.63	0.60	0.96	0.88	0.47	0.46
Частичное выздоровление	201	0.71	0.69	0.77	0.74	0.66	0.65
Полное выздоровление	108	0.89	0.88	0.98	0.99	0.81	0.79
Прогрессирующая болезнь	480	0.97	0.82	0.96	0.79	0.98	0.85
Среднее		0.8	0.75	0.92	0.85	0.73	0.69
Положительный ответ (полное или частичное выздоровление)	309	0.8	0.71	0.85	0.88	0.75	0.61

Объяснение причин

- Для поиска причин использовался GAAQ+JSM и AQ+JSM
- Примеры объяснений:
  - Летальный исход: «Кол-во клеток введенных за одну вакцинацию = высокое», «Индекс Карновского = низкий», «Возраст = выше среднего», «Возраст = выше среднего и пол = женский»
  - Стабильная или прогрессирующая болезнь: «Индекс Карновского = высокий», «Кол-во клеток введенных за одну вакцинацию = высокое», «DTH = не проводилось и Возраст = низкий», «Возраст = низкий и Лечение до иммунизации = химио-, гормональная, иммуно- и радиотерапии»
  - Частичное или полное выздоровление: «Кол-во клеток введенных за одну вакцинацию = высокое»

Boyko A. A., Kaidina A. M., Kim Y. C., Lupatov A. Yu., Panov A. I., Suvorov R. E., Shvets A. V. A Framework for Automated Meta-Analysis: Dendritic Cell Therapy Case Study // Труды конференции «IEEE Intelligent Systems IS'16». - София, Болгария, 2016. - сс. 160-166.

Проблема разметки данных

- Для обучения необходимо разметить много обучающих примеров
- Сложно настраиваться на новую предметную область, нужно снова размечать тексты
- Для новой предметной области заранее неизвестно, какие методы машинного обучения будут наиболее подходящими и какие признаки будут наиболее информативными

- Надо ускорить процесс разметки и обучаться сразу во время разметки

Активное онлайн обучение (Active On-Line Learning)

1. Эксперт размечает в текущем тексте несколько примеров целевой информации. Размеченные примеры становятся положительными обучающими примерами
2. Выполняется (до)обучение метода извлечения информации
3. Выполняется извлечение информации из текущего текста с помощью (до)обученного на шаге 2 метода извлечения
4. Эксперт оценивает результаты извлечения, отмечая правильно и неправильно извлеченные на шаге 3 примеры, которые полагаются положительными и, соответственно, отрицательными примерами. Кроме того, эксперт отвечает на вопросы относительно значимости признаков. Далее итерация повторяется, происходит переход на этап 1 или 2

- Процесс останавливается в любое время по желанию эксперта или при достижении требуемого уровня качества извлечения
- Результатом процесса является размеченный корпус и обученный метод извлечения информации

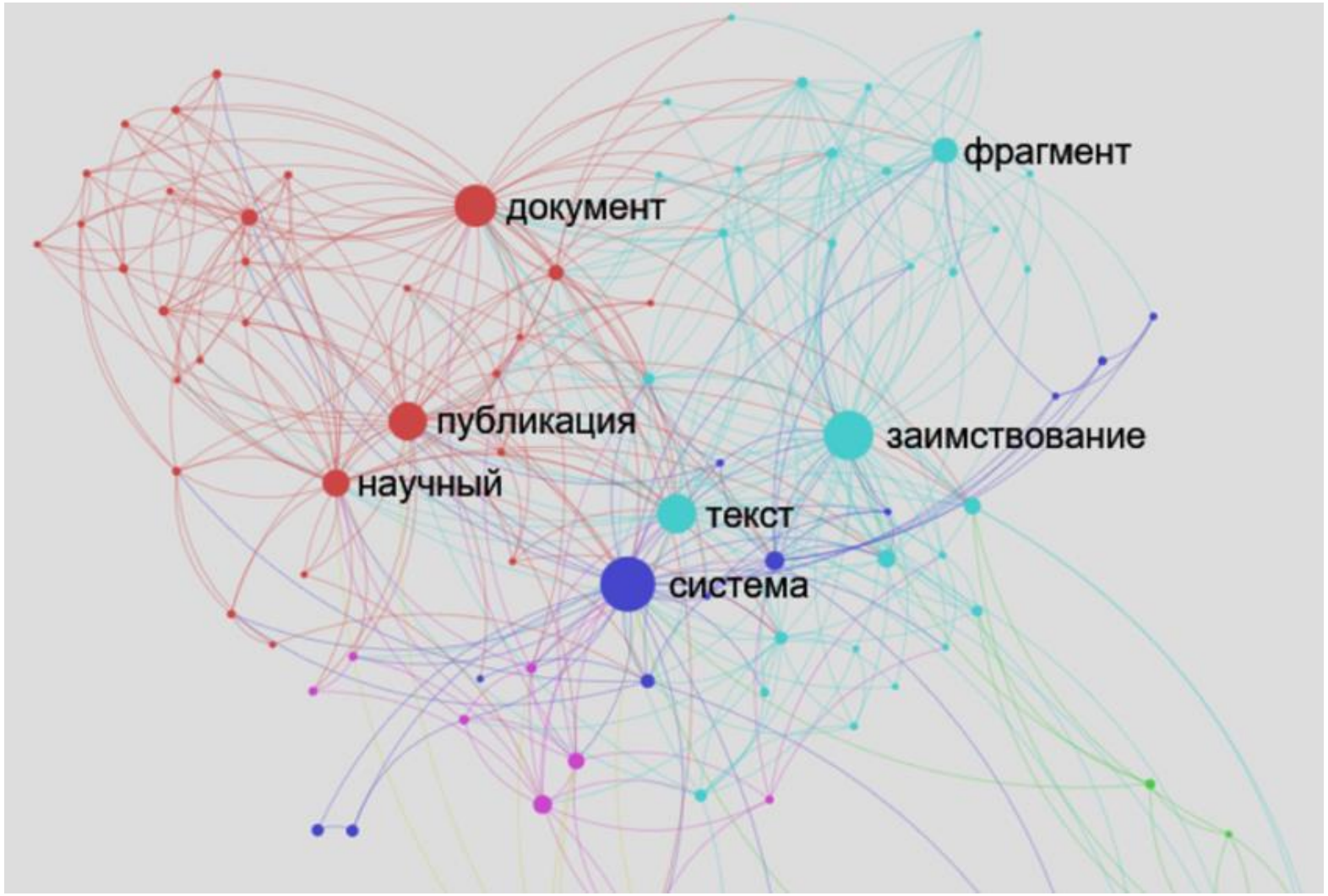
Особенности платформы

- Для начала обучения достаточно 2-3-х примеров
- Интерактивное взаимодействие системы с пользователем
- Пользователь может выбирать значимые по его мнению признаки
- Учитывается визуальное форматирование текста в документе (отступы, таблицы, размер шрифта и т.д.)
- Графовое представление всей лингвистической информации и аннотаций. Универсальность. Обучение происходит в графовом признаковом пространстве
- Интеграция различных методов машинного обучения

Что ещё извлекаем

- Чрезвычайные ситуации (взрывы, стихийные бедствия, техногенные катастрофы и т.п.)
- Финансовые ситуации (покупки-продажи, инвестиции, финансирование мероприятий, вложения в инфраструктуру и т.п.)

Извлечение знаний из больших массивов текстов



Открытое извлечение информации

- Получение из текстов полезной информации, структура и семантика которой заранее жестко не фиксированы.
- Основано на идее машинного обучения без учителя.
- Позволяет быстро решить задачи обработки текстовой информации без привлечения труда экспертов предметной области.

III. Анализ социальных медиа

Социальная напряженность

Задача

- Даны сообщения социальных медиа (СМ) – блогов, форумов, социальных сетей
- Необходимо проанализировать и сопоставить проявления социальной напряженности в реальности и в сетях в заданный период времени (протестные акции в России в 2011-2012 годах)

Проблема

- Как измерить социальную напряженность по сообщениям в социальных медиа?

- Предварительно сообщества разделены на напряженные (национализм, …) и нейтральные (кошки, …)

Маркеры социальной напряженности

Маркеры социальной напряженности – количественные показатели напряженности в активности интернет-сообщества

Типы маркеров:

- Маркеры активности
- Психолингвистические маркеры
- Лексические маркеры
- Семантические маркеры

Последние три основаны на анализе текстов сообщений СМ

Психолингвистические показатели эмоционального напряжения

Тексты, написанные здоровыми людьми в состоянии эмоционального напряжения, содержат индикаторы неблагополучия, которые отсутствуют в текстах тех же авторов, написанных в другое время (раньше и позже эмоциогенной ситуации)

Повышение или понижение выраженности соответствующих психолингвистических показателей отражает, таким образом, текущее эмоциональное состояние автора

Массовое повышение индикаторов эмоционального напряжения свидетельствует об эмоциональном заражении – запускаются процессы группирования на основе общности аффекта

Интернет сообщество в такие моменты своего существования может рассматриваться как потенциально обладающее готовностью к переходу к согласованным действиям в реальности

Психолингвистические маркеры

1. Количество слов в предложении.
2. Коэффициент определенности действия.
3. Количество глаголов в пассивном залоге.
4. Средняя длина слова.
5. Отношение количества инфинитивов к общему числу глаголов.
6. Количество безличных глаголов.
7. Количество местоимений.
8. Коэффициент Трейгера отношение количества глаголов к количеству прилагательных.
9. Количество глаголов несовершенного вида.
10. Количество местоимений 1-го лица множественного числа.
11. Количество инфинитивов.
12. Отношение числа глаголов и существительных к числу прилагательных и наречий.
13. Количество глаголов первого лица, единственного числа, прошедшего времени.
14. Количество предложений в тексте.
15. Средний размер предложения в словах.
16. Отношение количества глаголов будущего времени к общему количеству глаголов.
17. Количество местоимений 3-го лица множественного числа.

Словарные лексические маркеры

- Обозначение негативных эмоциональных и телесных состояний: *гнев противно беситься отвратительный бояться ужас охренеть …*
- Слова с деструктивной семантикой: значения разрушения, уничтожения, преобразующего действия на объект: *разрушить уничтожить исковеркать …*
- Лексика физического насилия: *бить ранить конфликт накостылять звезданути …*
- Инвективная лексика: *сволочь мразь шушера быдло выродок гадость кобель сука щенок козёл кобыла свинья дурак идиот имбецил кретин маразм …*
- Лексика протестного поведения: *агитация, анархия, баррикады, бунт….*
- Глаголы: *собираться, выходить, протестовать и т.п (300-400) глаголов.*
- Призывы: *пора, хватит терпеть, место сбора..., собираемся, все, кому небезразлична (судьба) и др. (несколько десятков)*

Экспериментальные лексические маркеры

Сайты с высокой выраженностью напряженности

- править
- заявить
- произойти
- состояться
- требовать
- поддерживать
- задержать
- начаться
- убивать
- выражать
- поддержать
- собирать
- убить
- призывать
- собраться
- бороться ….

Сайты без признаков напряженности

- любить
- хотеть
- понравиться
- ждать
- жить
- хотеться
- нравиться
- выглядеть
- работать
- поздравлять
- потрясать
- мечтать
- надеяться
- путешествовать
- простить
- учиться...

Семантические маркеры

- Выражают значения слов, например:
  - Деструктив – объект разрушающего воздействия (взорвать **дом**)
  - Директив – направление движения (выйти **на площадь**)
  - Ликвидатив – объект, прекращающий существование (убить **человека**)
  - Результатив –следствие (привести **к кризису**).
  - И др.
- Вычисляется количество значений в текстах сообщений, а также анализируются получившие значения слова
- Пример: слова со значением ликвидатив в период 7-15 декабря 2010 года (акция националистов на манежной площади): человек, парень, егор, ребенок, свиридов, большинство, кавказец, я, вы, павел …
- Пример: слова со значением ликвидатив в период 11-13 июня 2010 года (этнические беспорядки в Киргизии): узбек, человек, беспорядок, житель, мера, женщина, средство, большинство, брат, бандит …

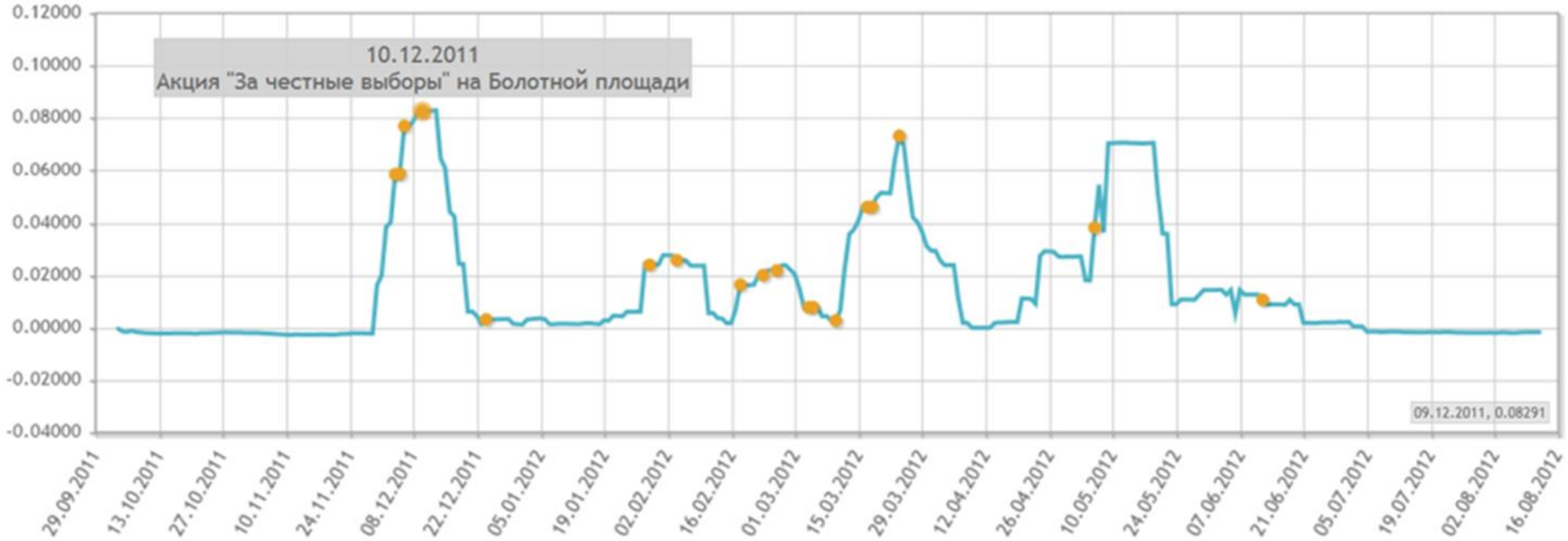
Эксперименты

- Коллекция нейтральных и стрессовых блогов Живого Журнала, отобранных экспертами
- 5 402 599 сообщений, оставленных за период с 1 января 2009 года по 15 августа 2012 года
- 64 маркера
- Анализ событий – Массовые митинги. 21 событие. Декабрь 2011 – Июль 2012

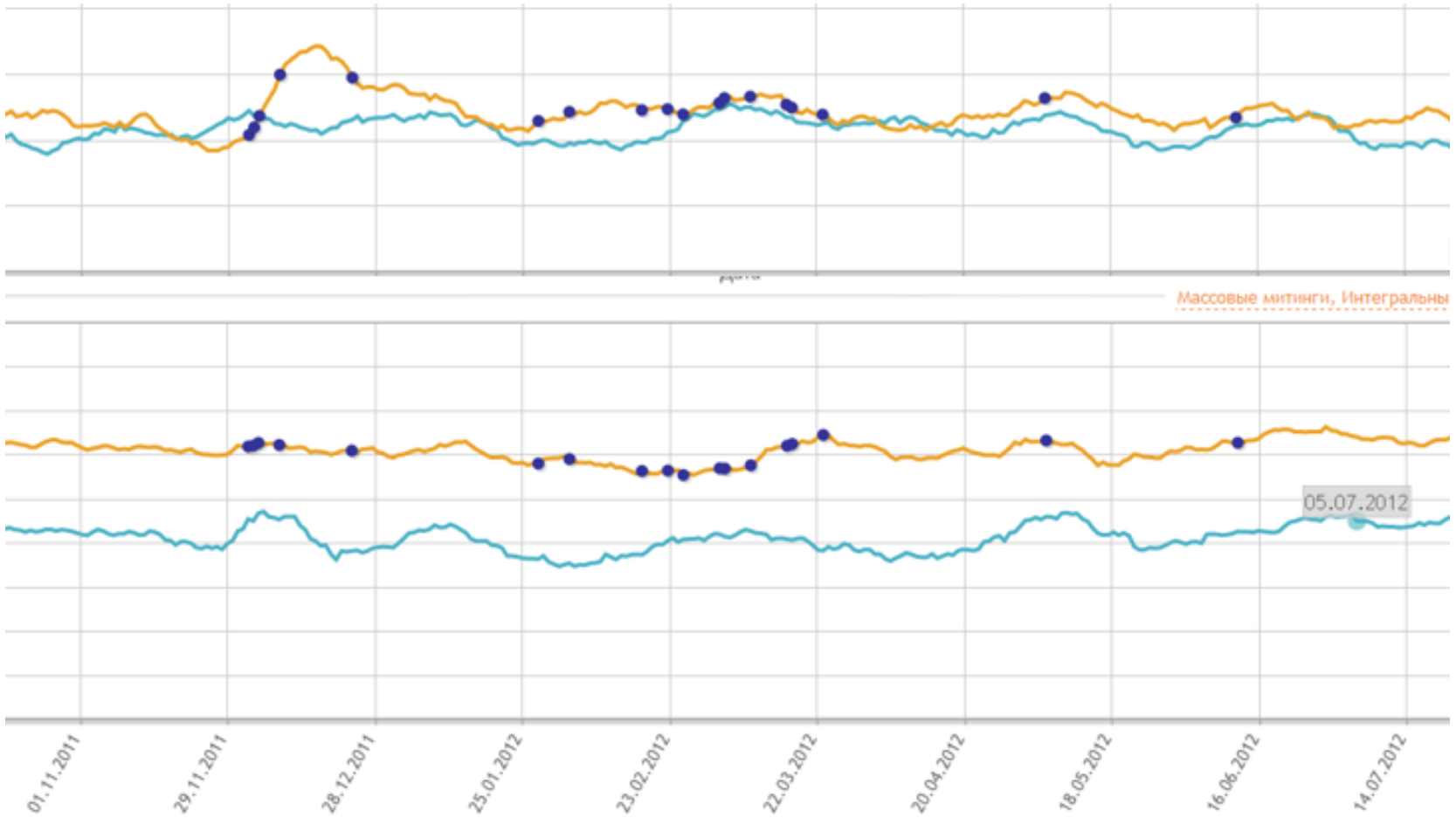
- Анализируется динамика отдельных маркеров и интегрального показателя напряженности

Динамика социальной напряженности

- Значения интегрального маркера напряженности



Сравнение степени напряженности в разных сообществах



III. Перспективы

Понимание текстов компьютером для человеко-машинного взаимодействия на естественном языке

- Интерпретация команд
- Ответы на вопросы
- Ведение диалога
- Автоматическое формирование баз знаний по массивам текстов и вывод в них
- Синтез связного текста по базам знаний
- Анализ и синтез речи

Создание интеллектуальных виртуальных ассистентов

- Ассистенты-консультанты по подбору товаров и услуг
- Ассистенты-эксперты, помогающие искать и анализировать информацию
- Ассистенты-собеседники, поддерживающие живую беседу
  - обладают картиной мира и эмоциями
  - воспринимают картины мира и эмоции собеседника
- Банковские ассистенты
- Медицинские ассистенты
- И др.

Благодарю за внимание!

Вопросы?

ivs@isa.ru

Наши основные публикации по теме

G. Osipov. Methods for Extracting Semantic Types of Natural Language Statements from Texts // 10th IEEE International Symposium on Intelligent Control 1995, Monterey, California, USA, Aug. 1995

Osipov G. S., Smirnov I.V., Tikhomirov I. A., Vybornova O.V, Zavjalova O. S. Linguistic Knowledge for Search Relevance Improvement // Proceedings of Joint conference on knowledge-based software engineering JCKBSE'06, IOS Press, 2006 - P. 294-302.

Смирнов И.В. Метод автоматического установления значений минимальных синтаксических единиц текста // Информационные технологии и вычислительные системы. – 2008. – №3. – С. 30-45.

Осипов Г.С., Смирнов И.В., Тихомиров И.А. Реляционно-ситуационный метод поиска и анализа текстов и его приложения // "Искусственный интеллект и принятие решений". – №2 – 2008. – С. 3-10.

Тихомиров И.А., Смирнов И.В. Применение методов лингвистической семантики и машинного обучения для повышения точности и полноты поиска в поисковой машине Exactus. //Труды международной конференции Диалог'2009. - С. 483-487.

Olga Vybornova, Ivan Smirnov, Ilya Sochenkov, Alexander Kiselyov, Ilya Tikhomirov, Natalya Chudova, Yulia Kuznetsova and Gennady Osipov. Social Tension Detection and Intention Recognition Using Natural Language Semantic Analysis (on the material of Russian-speaking social networks and web forums) // Proceedings of the European Intelligence and Security Informatics Conference (EISIC) 2011, p. 277-281, September 12-14, 2011 Athens, Greece.

Gennady Osipov, Ivan Smirnov, Ilya Tikhomirov, Artem Shelmanov Relational-Situational Method for Intelligent Search and Analysis of Scientific Publications // Proceedings of the Integrating IR technologies for Professional Search Workshop, Moscow, Russia, 24-March-2013, pp. 57-64.

И.В. Смирнов, А.О. Шелманов, Е.С. Кузнецова, И.В. Храмоин Семантико-синтаксический анализ естественных языков. Часть II. Метод семантико-синтаксического анализа текстов // Искусственный интеллект и принятие решений. М.: ИСА РАН – 2014. – №1 – С. 11-24.

Наши основные публикации по теме

Shelmanov A. O., Smirnov I. V., Methods for Semantic Role Labeling of Russian Texts // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2014). Issue 13 (20). – 2014. – pp. 580-592.

Shelmanov A. O., Smirnov I. V., Vishneva E. A. Information extraction from clinical texts in Russian // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2015). Issue 14 (21). – 2015. – V1. – pp. 537-549.

Roman Suvorov, Ivan Smirnov, Konstantin Popov, Nikolay Yarygin, Konstantin Yarygin. Assessment of the Extent of the Necessary Clinical Testing of New Biotechnological Products Based on the Analysis of Scientific Publications and Clinical Trials Reports // Proceedings of the International Conference on Pattern Recognition Applications and Methods. - Scitepress. - 2015. – Vol. 2. - pp. 343-348.

Zubarev, D., Sochenkov, I.: Using Sentence Similarity Measure for Plagiarism Source Retrieval — Notebook for PAN at CLEF 2014. In: CEUR Workshop Proceedings, CEUR-WS.org, Eds. L. Cappellato, N. Ferro, M. Halvey and W. Kraaij. 2014. P.p. 1027–1034

Sochenkov, Ilya, Denis Zubarev, Ilya Tikhomirov, Ivan Smirnov, Artem Shelmanov, Roman Suvorov, and Gennady Osipov. "Exactus Like: Plagiarism Detection in Scientific Texts." In Advances in Information Retrieval, pp. 837-840. Springer International Publishing, 2016.

А.А. Баранов, Л.С. Намазова-Баранова, И.В. Смирнов, Д.А. Девяткин, А.О. Шелманов, Е.А. Вишнева, Е.В. Антонова, В.И. Смирнов. Технологии комплексного интеллектуального анализа клинических данных // Вестник РАМН, 2016, №2. – С.160-171.