

Введение в КЛ

Занятие 1

КЛ и NLP

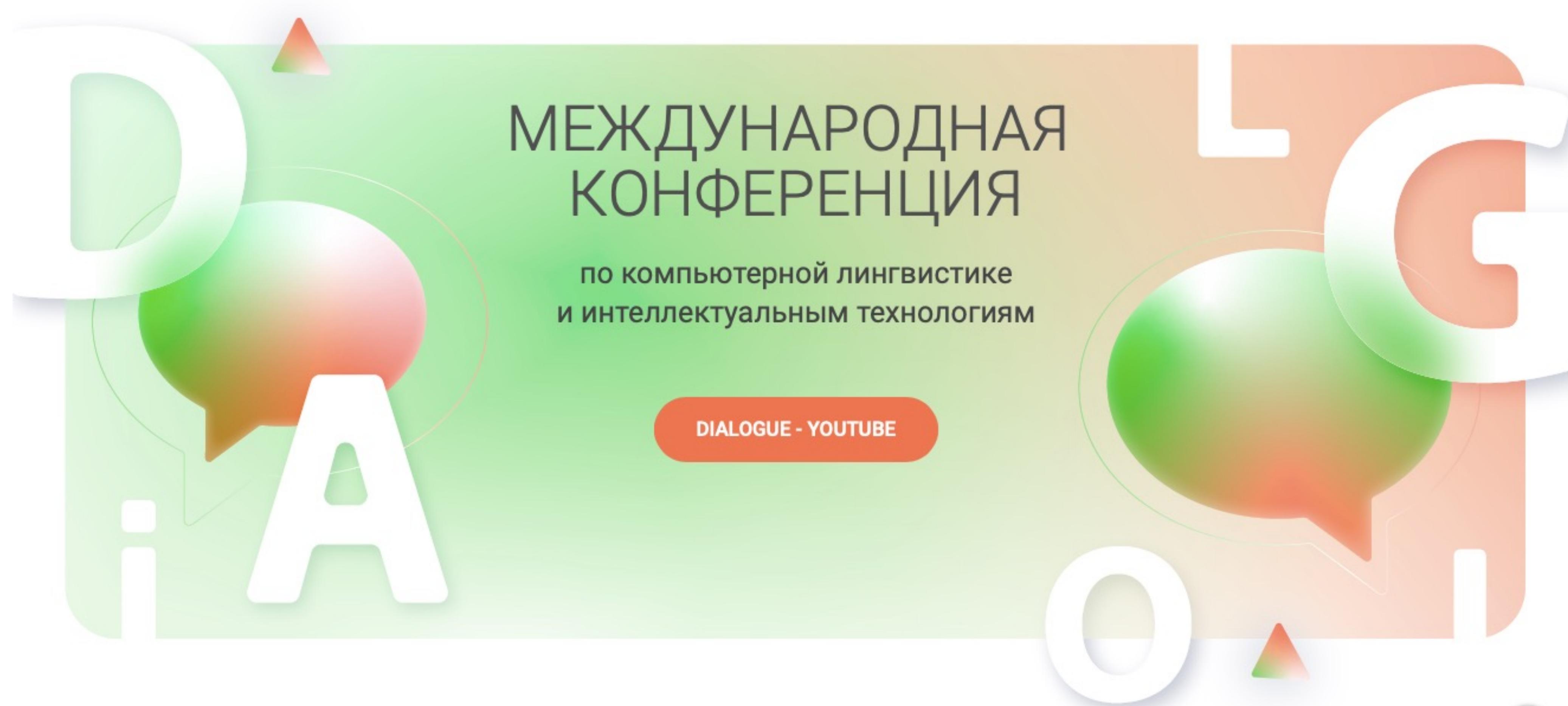
Разное или одинаковое

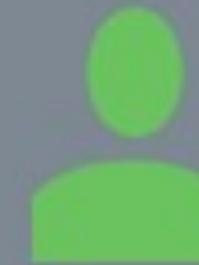
В.П. Селегей:

- **КЛ** - ближе к лингвистике. Комп. поддержка языковых исследований. Цели как у Теоретической Лингвистики: полно описать структуру и функции, объяснить явления ЕЯ.
- **NLP (Natural Language Processing)** - «создание лингвистических пылесосов». Область решения практических задач, связанных с языком. Раздел AI.
- Вышли из одной (примерно) точки и все больше расходятся.

Конференция Dialogue

КЛ и NLP





300

ЧЕЛОВЕК В СРЕДНЕМ ПОСЕЩАЮТ
КОНФЕРЕНЦИЮ ЕЖЕГОДНО



2499

ДОКЛАДОВ СДЕЛАНО ЗА ВСЕ
ВРЕМЯ



27

СОРЕВНОВАНИЙ ПРОВЕДЕНО
В РАМКАХ DIALOGUE EVALUATION

«Диалог» является крупнейшей в России конференцией по компьютерной лингвистике, служит уникальным мировым форумом для обсуждения методов компьютерного анализа русского языка. Конференция позволяет оценить уровень российской компьютерной лингвистики и определить векторы её развития.

Ежегодно конференция собирает ведущих лингвистов, инженеров, специалистов в области автоматической обработки языка и представителей бизнеса.

Доклады принимаются по результатам рецензирования и публикуются в сборнике «Компьютерная лингвистика и интеллектуальные технологии».

Итоги соревнований

2021

Упрощение текстов

Семантические сдвиги

Малоресурсные языки

Кластеризация

Семантические скетчи

Нормализация текстов

2020

Таксономия

GramEval

RuREBus

2019

Генерация заголовков

Гэппинг

Малоресурсные языки

Разрешение анафоры

2018

Разрешение лексической
многозначности

2017

Поиск заимствований
Морфологический анализ

2016

Анализ тональности
Выделение сущностей
Исправление опечаток

Evaluation проводят с 2010 года...

КЛ и NLP

2015

Анализ тональности

Семантическая близость

2014

Разрешение анафоры

2013

Анализ тональности

Машинный перевод

2012

Анализ тональности

Синтаксис

2010

Морфология

Evaluation проводят с 2010 года...

КЛ и NLP

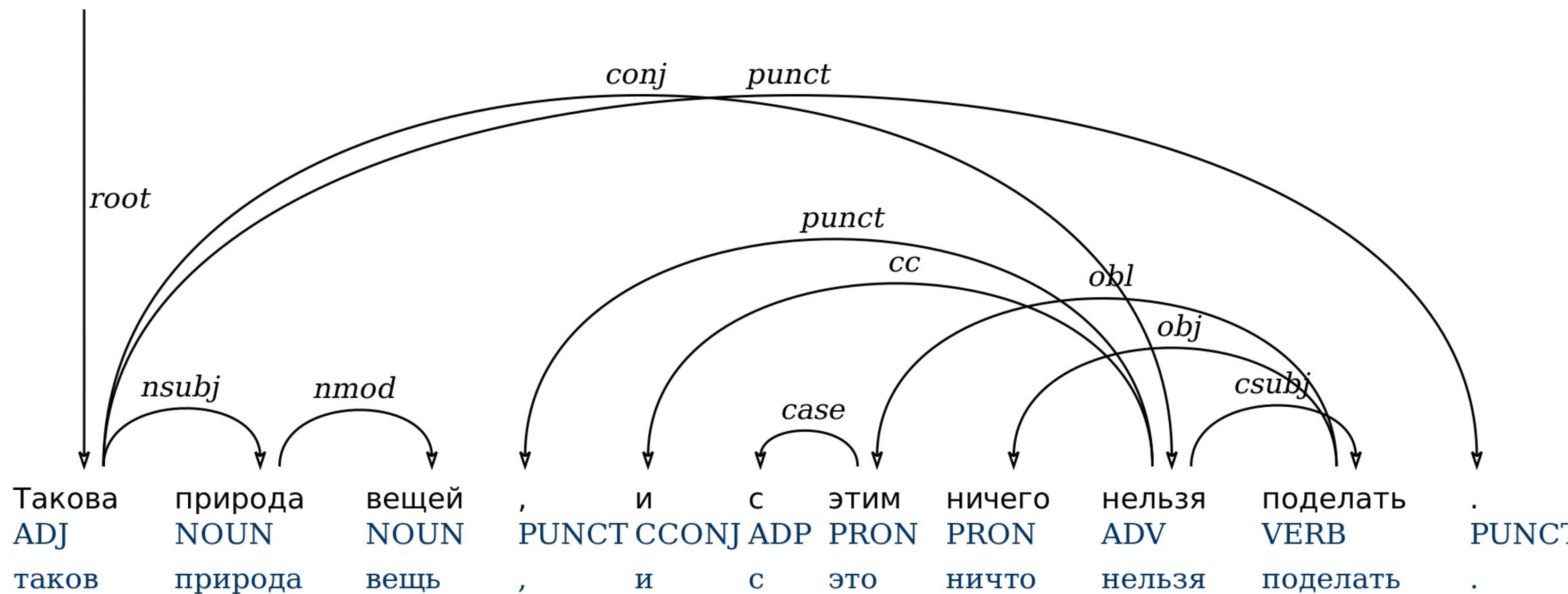
Почему соревнование - это важно?

- Установить стандарты разметки
- Установить метрики оценки качества (evaluation)
- Золотая разметка (gold set)
- Baseline (база, что пытаемся опередить)
- SOTA (State of the Art - текущее положение дел)

Кл. Задачи

1. Лингвистический анализ текста:

- Морфологический парсинг (try it!)
- Синтаксический парсинг



Кл. Задачи

1. Лингвистический анализ текста:

- Семантический анализ

Семантический скетч для глагола "выходить:TO_TAKE_PLACE":

Ch_Relation_Coincidence	Modality	Object_Situation	Ch_Evaluation	Time	Locative
наоборот	само_собой	заминка	складно	четверть	сумма
вышло наоборот	вышло само собой	вышла заминка	выходило складно	выходила за четверть	вышло в сумме
такой	так	скандалить	некоторый	как-то	рассказ
вышло так	вышло так	вышел скандал	вышло некрасиво	вышло как-то	выходит в рассказах гуцко
иной	на_деле	размолвка	скверный	восьмой	практик
вышло по-иному	вышло на деле	вышла размолвка	выходило скверно	вышла восьмого марта	выходит на практике
похожий	неправдоподобный	казус	красивый	иной_раз	конкурентка
выходило очень похоже	вышло крайне неправдоподобно	вышел казус	вышло красиво	выходит иной раз	выходит у конкурентки
другой	криво	неприятность	ничего	подчас	Земский
вышло по-другому	вышло криво	вышли неприятности	вышло ничего	выходило так подчас	выходит у него
этакий	кривовато	ссориться	паскудный	ноябрь	Вяльцев
вышло этак	выходило кривовато	вышла ссора	вышло паскудно	выходило на прошлогодний ноябрь	выходит у вяльцева

Кл. Задачи

Семантический скетч для глагола "выходить: TO_FRONT":

Locative_FinalPoint	Object	Locative_PartAsOrientation	Locative_Orientation_FinalPoint	Time	Locative
двор	окно	окно	запад	поныне	Соня
выходили во двор	выходили окна	выходила окнами	выходили на запад	выходит и поныне	выходит у нас
сад	балкон	дверь	север	частенький	ярус
выходили в сад	выходил балкон	выходили дверями	выходят на север	выходил частенько	выходили на втором ярусе
дворик	фасад	фасад	восток	параллельный	n_этажка
выходило во внутренний дворик	выходил его главный фасад	выходили фасадом на тверскую	выходит		
улица	окошко	конец			
выходили на улицу	выходило окошко	выходит одним концом			
проспект	веранда	стена			
выходили на проспект	выходила веранда	выходит северной стеной			
север	подъезд		на		
выходили на север	выходили на подъезды		выходили напр		

WORD SKETCH

выйти as verb 2,861,777x

↔	≡≡	□	×	↔	≡≡	□	×	↔	≡≡	□	×	↔	≡≡	□	×	↔	≡≡	□	×	↔	≡≡	□	×
subject				post_prep				pp_на				pp_из				pp_в				adv_modifier			
книга	...	из	...	на	...	за	...	улица	...	на	...	около	...	пенсия	...	комната	...	строй	...	строй	...	финаль	...
вышла книга		вышел из		вышел на		вышла за		вышел на улицу		вышел на		вышел около		вышел на пенсию		вышел из комнаты		вышел из строя		вышел из строя		вышли в финал	...
версия	...	на	...	на	...	за	...	сцена	...	на	...	на	...	пенсия	...	комната	...	мода	...	мода	...	полуфинал	...
Вышла новая версия		вышел на		вышел на		вышла за		вышел на сцену		вышел на		вышел на		вышел на пенсию		вышел из комнаты		вышел из моды		вышел из моды		вышли в полуфинал	...
постановление	...	за	...	за	...	около	...	крыльцо	...	за	...	около	...	балкон	...	комната	...	крыльцо	...	балкон	...	отставка	...
вышло постановление		вышел за		вышел на		вышел около		вышел на крыльцо		вышел за		вышел около		вышел на балкон		вышел из комнаты		вышел из строя		вышел из строя		вышли в отставку	...
фильм	...	около	...	крыльцо	...	балкон	...	тюрьма	...	около	...	балкон	...	тюрьма	...	комната	...	тюрьма	...	балкон	...	эфир	...
ошибочка	...	вышел около		вышел на		вышел на		ванная	...	вышел около		вышел на		вышел из тюрьмы		вышел из комнаты		вышел из строя		вышел из строя		вышли в эфир	...
ошибочка вышла		ошибочка вышла		ошибочка вышла		ошибочка вышла		ринг	...	ошибочка вышла		ошибочка вышла		ошибочка вышла		комната		ошибочка вышла		ошибочка вышла		ошибочка вышла	...
альбом	...	альбом вышел		альбом вышел		альбом вышел		через	...	альбом вышел		альбом вышел		альбом вышел		балкон		альбом вышел		альбом вышел		альбом вышел	...
								через	...	альбом вышел		альбом вышел		альбом вышел		балкон		альбом вышел		альбом вышел		альбом вышел	...
указ	...	вышел указ		вышел указ		вышел указ		экран	...	вышел указ		вышел указ		вышел указ		балкон		вышел экран		вышел экран		вышел экран	...
								экран	...	вышел указ		вышел указ		вышел указ		балкон		вышел экран		вышел экран		вышел экран	...
издание	...	издание вышло в		издание вышло в		издание вышло в		к	...	издание вышло в		издание вышло в		издание вышло в		балкон		вышел экран		вышел экран		вышел экран	...
								к	...	издание вышло в		издание вышло в		издание вышло в		балкон		вышел экран		вышел экран		вышел экран	...
девушка	...	девушка вышла		девушка вышла		девушка вышла		в	...	девушка вышла		девушка вышла		девушка вышла		старт	...	девушка вышла		девушка вышла		девушка вышла	...
								в	...	девушка вышла		девушка вышла		девушка вышла		старт		девушка вышла		девушка вышла		девушка вышла	...
ко	...	ко		ко		ко		ко	...	ко		ко		ко		орбита	...	ко		ко		ко	...
																орбита	...					орбита	...

WORD SKETCH
Russian Web 2011 (ruTenTen11)
🔍
⬇️
👁️
🔍
 ⓘ
✖️

🔍
⬇️
👁️
🔍
 ⓘ
✖️

Кл. Задачи

1. Лингвистический анализ текста:

- Разрешение анафоры и кореференции

3. Великолепная «[Школа] злословия» вернулась в эфир после летних каникул в новом
формате.

4. В истории программы это уже не первый « ребрендинг ».

5. Сейчас с трудом можно припомнить, что начиналась «[Школа]...» на [канале] «Культура»
как стандартное ток[о]шоу, которое отличалось от других «кухонными» обсуждениями гостя,
что называется – «за глаза», и неожиданными персонами в качестве ведущих.

6. Писательница Татьяна Толстая и сценаристка Дуня Смирнова вроде бы не вполне
соответствовали принятым на российском телевидении стандартам телеведущих.

7. Впрочем, на [канале] «Культура» в роли телеведущих выступают и писатели, и
композиторы, так что в этом ничего сверхъестественного не было, а идея кухонных
обсуждений не слишком прижилась, и некоторые выпуски программы обходились
практически без них.

Кл. Задачи

Повторим

1. Лингвистический анализ текста:

- Морфологический парсинг
- Синтаксический парсинг
- Семантический анализ
- Разрешение анафоры и кореференции

Кл. Задачи

2. Жанровая классификация

Основная интерпретация класса
Аргументативные тексты
Личные блоги
Новостные тексты
Юридические тексты
Рекламные тексты
Научные тексты
Энциклопедические тексты
Инструкции
Художественная литература

Отбор языковых признаков

Языковые признаки

D. Biber

Agentless passives
Attributive adjectives
Average word length
By-passives
Demonstratives
First person pronoun
Indefinite pronoun
Independent clause coordination
Necessity modals
Nominalizations
Past tense
Perfect aspect
Phrasal coordination
Total prepositional phrases
Place adverbs
Possibility modals
Predicative adjectives
Private verbs
Public verbs
Second person pronoun
Split infinitives
Suasive verbs
That verb complements
Total other nouns
Wh-clauses
...



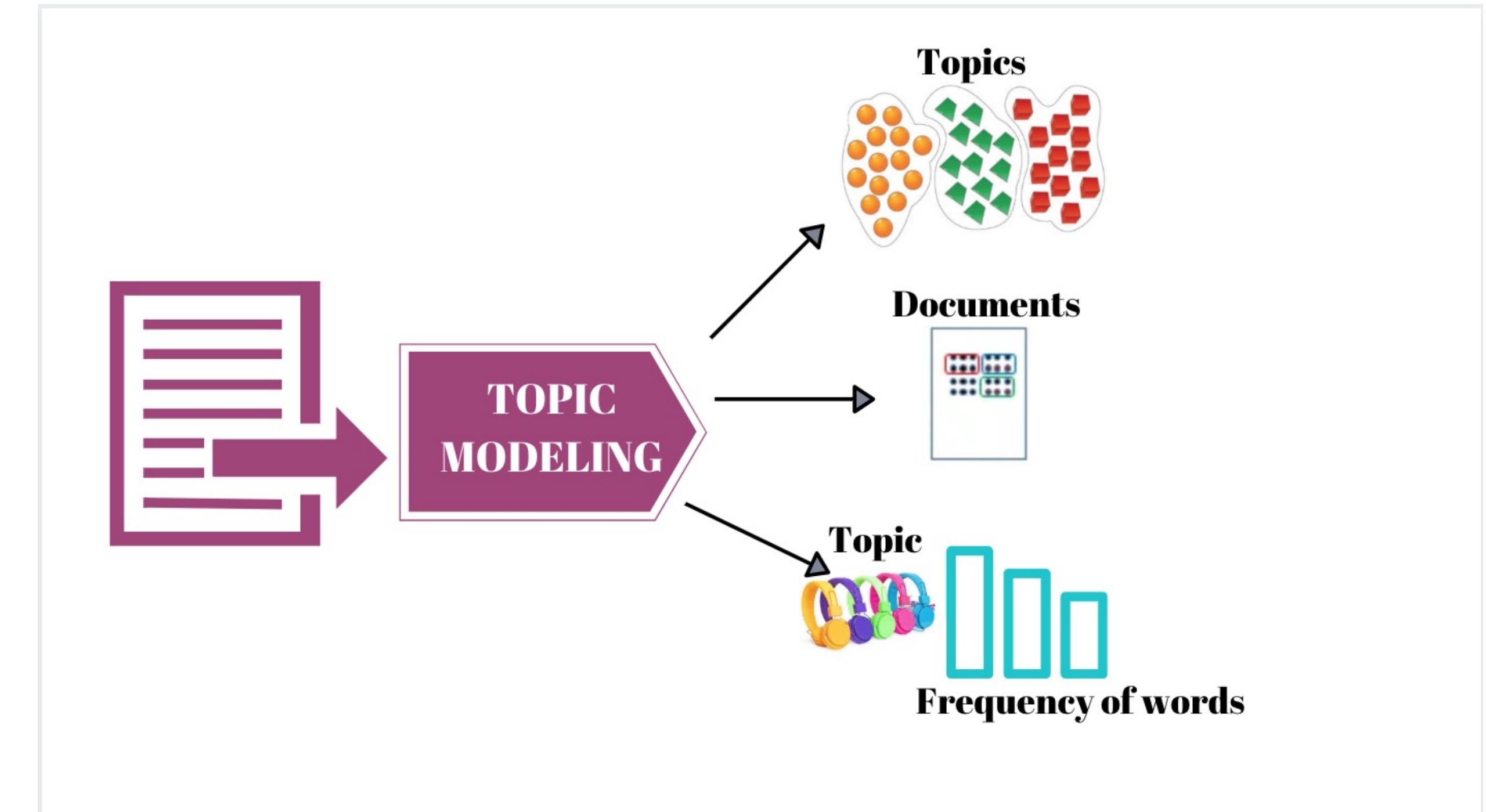
MDRus Analyser

Предложения с пассивными конструкциями без выраженного агента
Адъективное определение
Средняя длина слова
Предложения с пассивными конструкциями с выраженным агентом
Указательные местоимения *это, этот, тот*, в качестве определителей именных групп
Местоимения первого лица
Неопределенные местоимения
Сочинение клауз
Модальные слова со значением необходимости
Отлагольные существительные
Прошедшее время
Совершенный вид
Сочинение фразовых категорий
Указательные местоимения *это, этот, тот*, в качестве определителей именных групп
Наречия места
Модальные слова со значением возможности
Прилагательное в предикативной функции
Ментальные глаголы
Глаголы речи
Местоимения второго лица
Инфинитив
Изъяснительные придаточные
Все существительные
Относительные придаточные клаузы
...



Кл. Задачи

3. Тематическая
классификация/кластеризация
(тематическое моделирование)

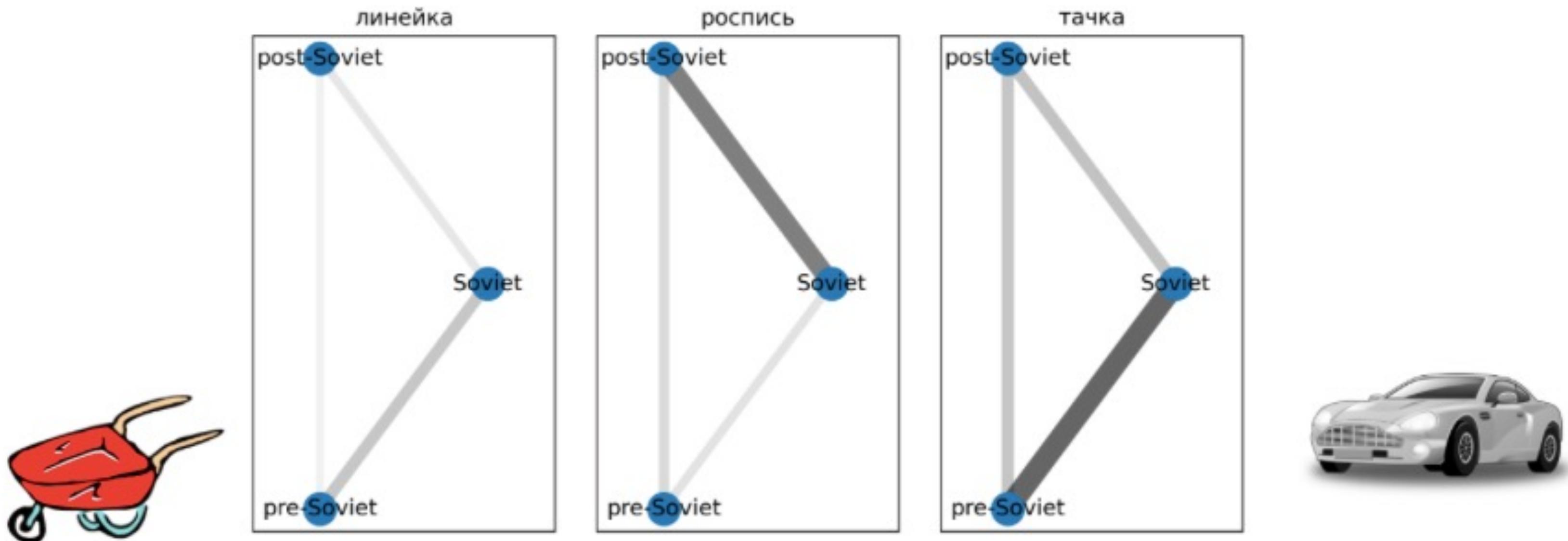


КЛ

Задачи

4. Анализ сложности

5. Диахронические сдвиги (RuShiftEval2021 - семантические сдвиги)



КЛ

Задачи

6. Разработка стандартов разметки всех видов анализов

снимаем	Vmip1р-а-е	снимать
шумовкой	Ncf sin	шумовка
,	,	,
чтоб	с	чтоб
масло	Ncnsnn	масло

стекло	Vmis-sna-p	стечь
.	SENT	.

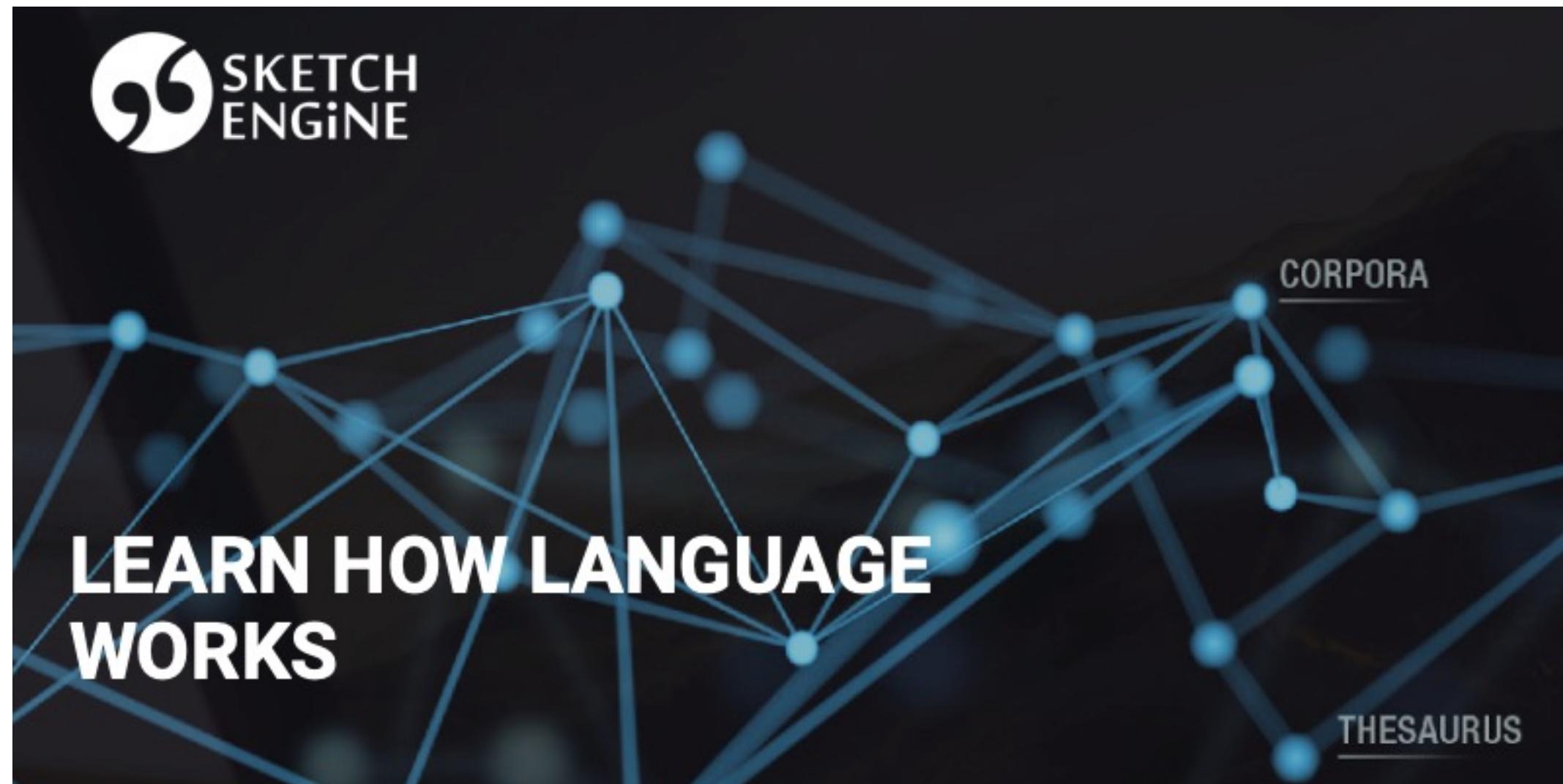
```
>>> from russian_tagsets import converters
>>> to_aot = converters.converter('opencorpora-int', 'aot')
>>> to_aot("NOUN,anim,masc plur,nomn")
С, од, мр, мн, им
```

- OpenCorpora (в.т.ч. русские словари [pymorphy2](#));
- aot.ru (в.т.ч. [pymorphy 0.5.6](#));
- Диалог-2010;
- A Positional Tagset for Russian (Jirka Hana and Anna Feldman, 2010);
- НКРЯ;
- Universal Dependencies (v1.4, v2.0, [Dialog-2017](#));

КЛ

Задачи

7. Создание корпусов



ГИКРЯ Меню

Настройка поиска

Лексико-грамматический поиск Экстралингвистические атрибуты

Добавить запрос

Искать среди слов: или найти не менее чем X результатов: 100

Имя запроса: Сохранить Удалить Поделиться с

Стат. запрос Включить дедупликацию

Включить?	Сегмент	Слов:	Текстов:
<input checked="" type="radio"/>	Вконтакте - 2014-2017	5115 млн.	191 млн.
<input type="radio"/>	Журнальный Зал - 1990-2018	320 млн.	73 тыс.
<input type="radio"/>	Живой Журнал - 2008-2020	17293 млн.	418 млн.

Запустить Остановить

Taiga Corpus

An open-source corpus for machine learning.

Home News Corpus Annotation Segments Downloads View on GitHub

КЛ

Задачи

1. Лингвистический анализ текста
2. Жанровая классификация
3. Тематическое моделирование
4. Анализ сложности
5. Диахронические сдвиги
6. Форматы разметки
7. Создание корпусов

NLP

Задачи

1. Машинный перевод

Русский ↔ Английский

Аллах велик × Allah is a bicycle
Allakh velik

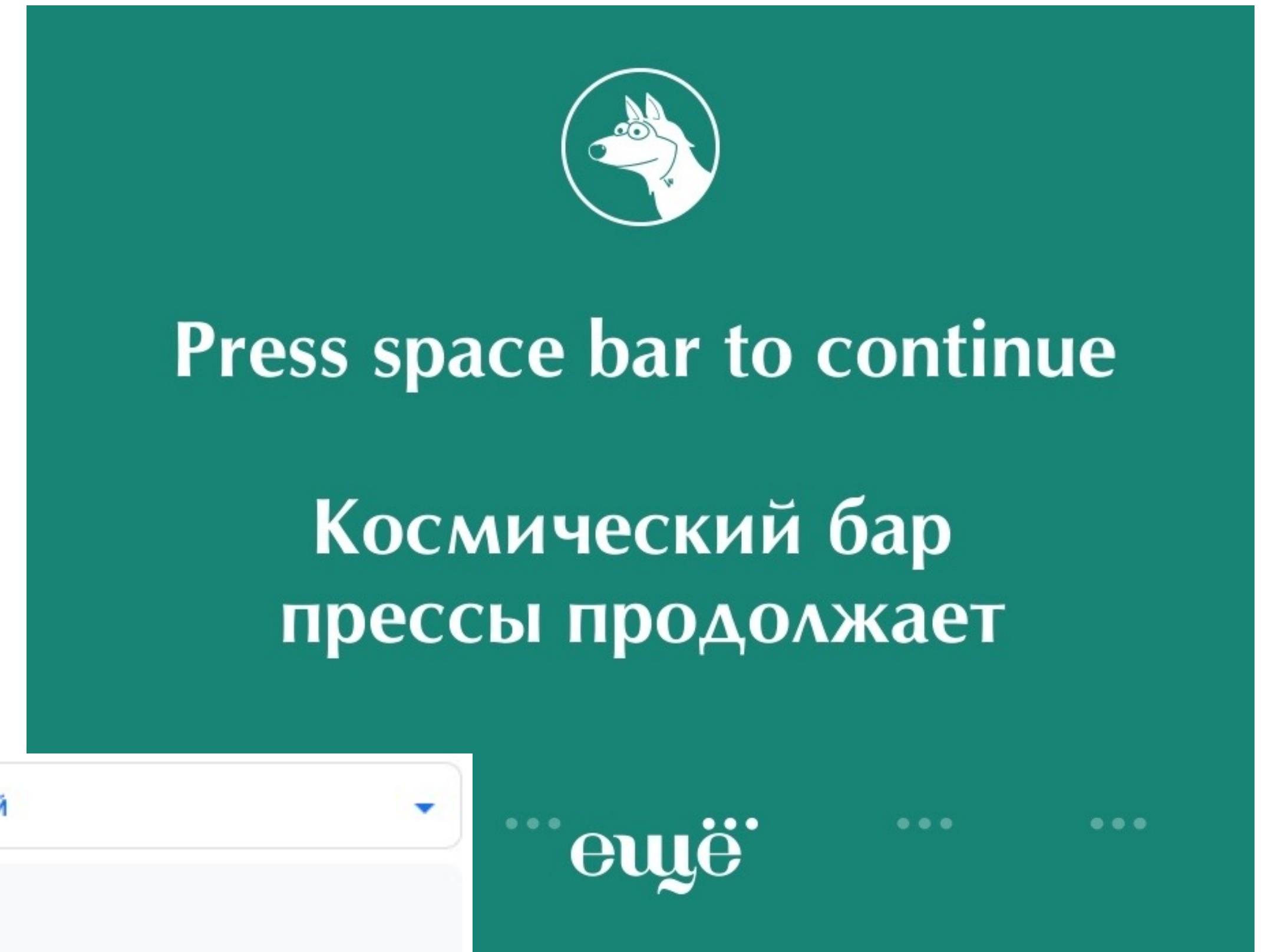
Английский ↔ Немецкий

thirty five × 35

...

ещё

Проверено



NLP

Задачи

2. Извлечение информации (Information extraction):

- NER - извлечение именованных сущностей
- разрешение анафоры
- выделение терминологии
- извлечение отношений

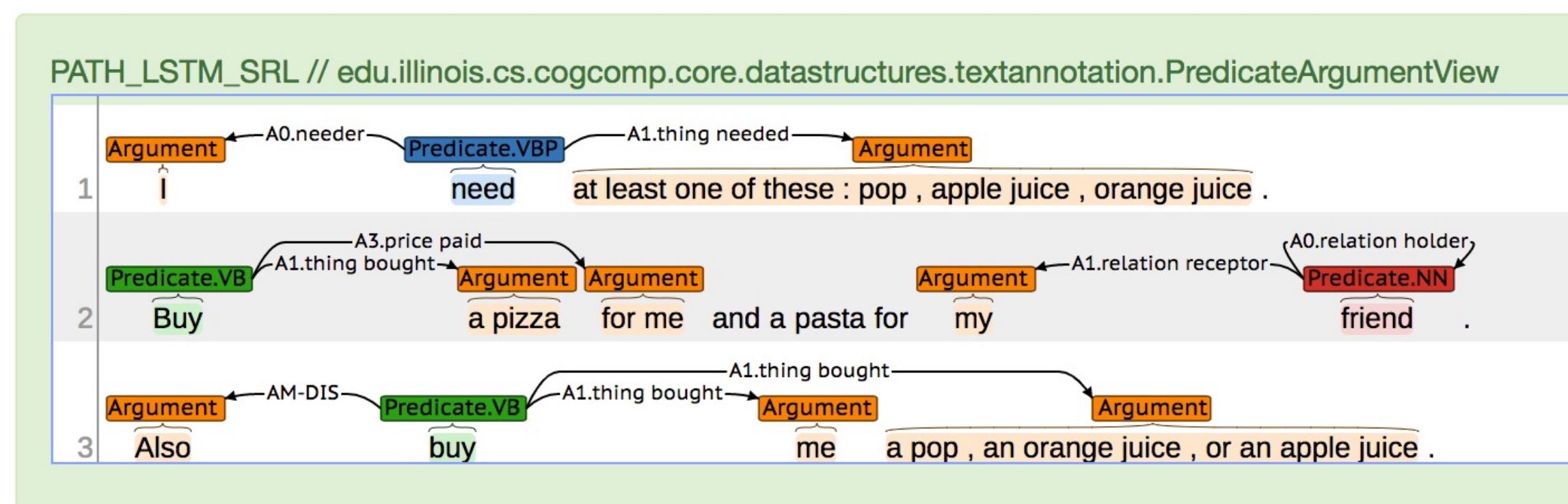


Figure 1: An example of NER application on an example text

In fact, the Chinese NORP market has the three CARDINAL most influential names of the retail and tech space – Alibaba GPE , Baidu ORG , and Tencent PERSON (collectively touted as BAT ORG), and is betting big in the global AI GPE in retail industry space . The three CARDINAL giants which are claimed to have a cut-throat competition with the U.S. GPE (in terms of resources and capital) are positioning themselves to become the future AI PERSON platforms . The trio is also expanding in other Asian NORP countries and investing heavily in the U.S. GPE based AI GPE startups to leverage the power of AI GPE . Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing one CARDINAL , with an anticipated CAGR PERSON of 45% PERCENT over 2018 - 2024 DATE .

To further elaborate on the geographical trends, North America LOC has procured more than 50% PERCENT of the global share in 2017 DATE and has been leading the regional landscape of AI GPE in the retail market. The U.S. GPE has a significant credit in the regional trends with over 65% PERCENT of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as Google ORG , IBM ORG , and Microsoft ORG .

NLP

Задачи

NER - try it!

почитать об этом решении: <https://natasha.github.io/ner/>

```
1 from navec import Navec
2 from slovnet import NER
3 from ipymarkup import show_span_ascii_markup as show_markup
4
5 text = '''
6 "Мы прекрасно помним, как все это было, наши граждане фактически были заманены
7 Эпизод с так называемыми вагнеровцами ФСБ считает уголовно наказуемым деянием
8 '''
9
10 navec = Navec.load('Downloads/navec_news_v1_1B_250K_300d_100q.tar')
11 ner = NER.load('Downloads/slovnet_ner_news_v1.tar')
12 ner.navec(navec)
13
14 markup = ner(text)
15 show_markup(markup.text, markup.spans)
```

"Мы прекрасно помним, как все это было, наши граждане фактически были
заманены ГУР и СБУ под руководством ЦРУ США в Беларусь... Цель была

ORG ORG ORG LOC LOC

известна, был вброс в КГБ республики о том, что эти люди приехали для

ORG

участия в массовых беспорядках на стороне оппозиции, белорусские
партнеры поэтому так и среагировали. Потом все стало на свои места,
несмотря на те попытки, которые предпринимала украинская сторона, мы
прекрасно помним звонок Зеленского президенту Лукашенко. Время все

PER PER

расставило по своим местам", – добавил он.

Эпизод с так называемыми вагнеровцами ФСБ считает уголовно наказуемым

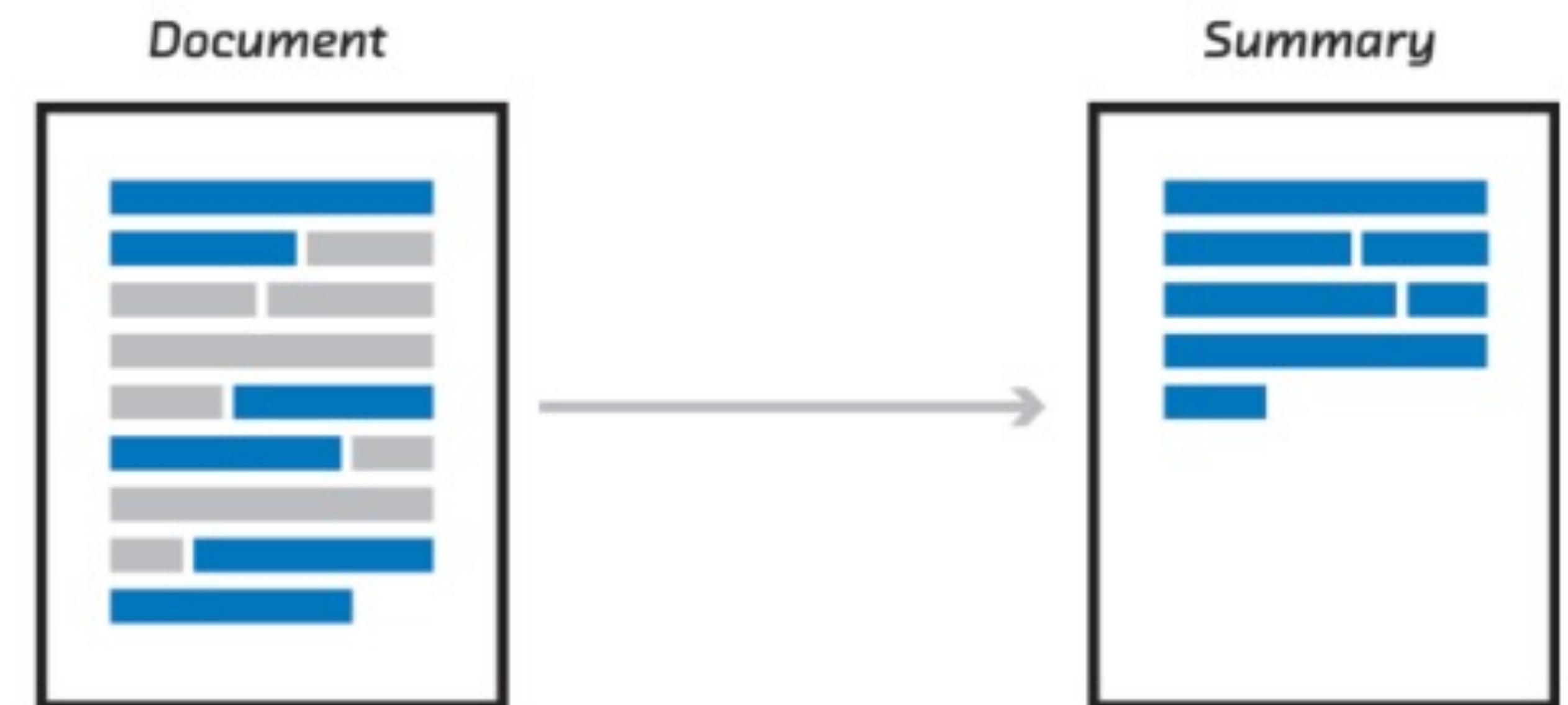
ORG

NLP

Задачи

3. Саммаризация / автоматическое реферирование

- аннотация к научным статьям
- реферирование новостей
- сниппеты сайтов
- резюме встреч



NLP

Задачи

4. Упрощение (Simplification)

There are 8 hard words. Learn 5 of them, or all of them? Hand-p
Reading time: 27 seconds. | Total points: 0 | ? | X

Rewordified text

Stats

Share

Print / Learning activities

Parts of speech



Navy Lieutenant Nunn arrived this morning at Lord Dartmouth's with letters from General Gage, Lord Percy, and Lieutenant-Colonel Smith concerning the events that occurred on April 19th and involved the British troops, the colonial militia, and several rebel parties. Lieutenant-Colonel Smith was concerned about the chaos that was occurring north of Boston and sent six light infantry companies to secure two bridges near Concord. On their way to complete this order, the companies found a group of armed country people near the road in Lexington. As the British troops approached them to find out why they had assembled, the members of this militia ran off in great confusion. During this time, several shots were fired at the British troops from behind a stone wall, a meeting house, and other homes. One British soldier

concerning the about the
occurred happened
militia group of armed citizens
several (more than two, but not a lot of)
concerned about worried about
chaos noise and confusion
occurring happening
assembled got together

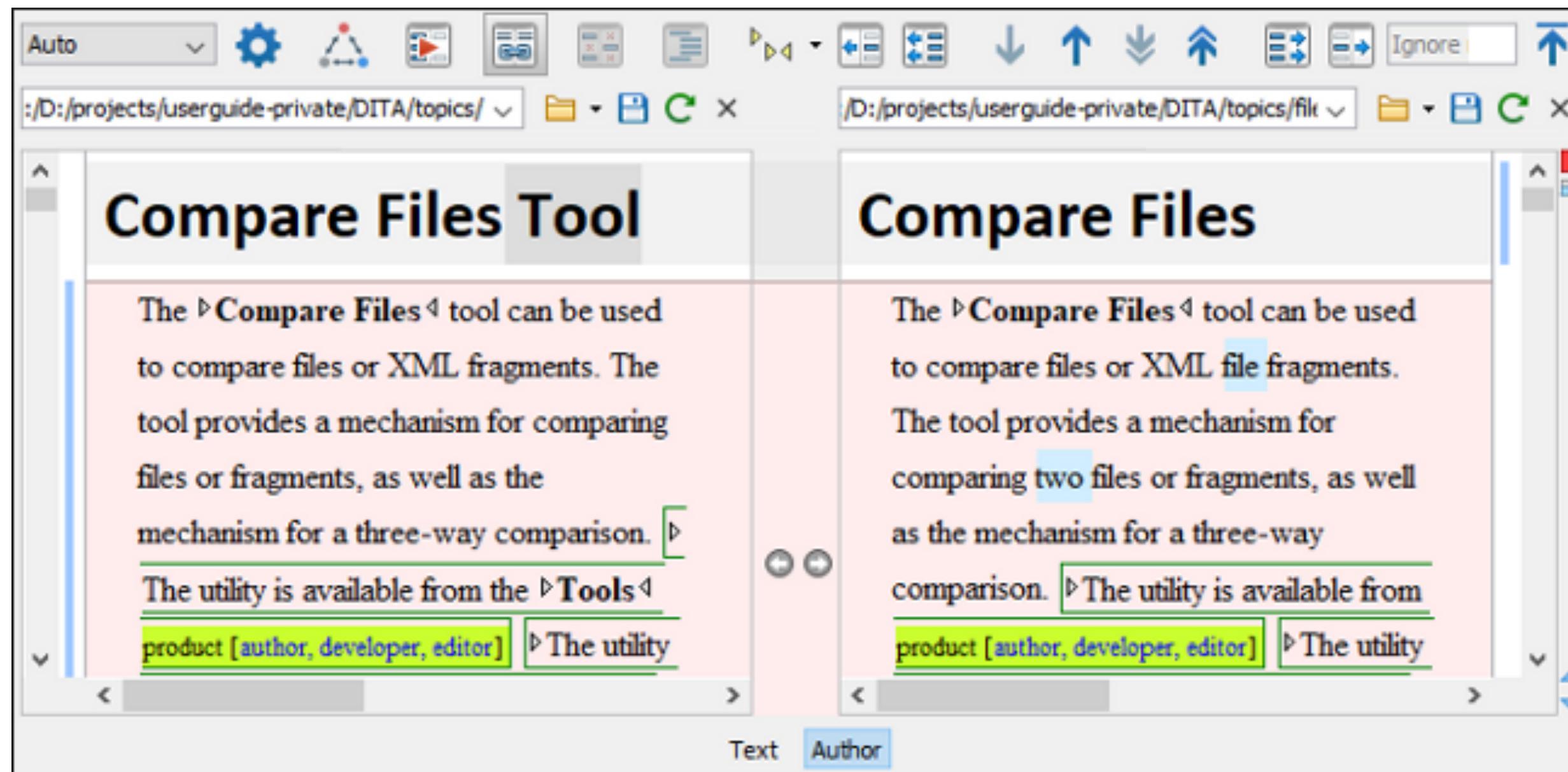
Example of a Complex Sentence	Example of a Simplified Sentence
Grammarly provides assistance in order to optimize users' communication.	Grammarly helps people communicate.

NLP

Задачи

5. Автоматическая классификация документов

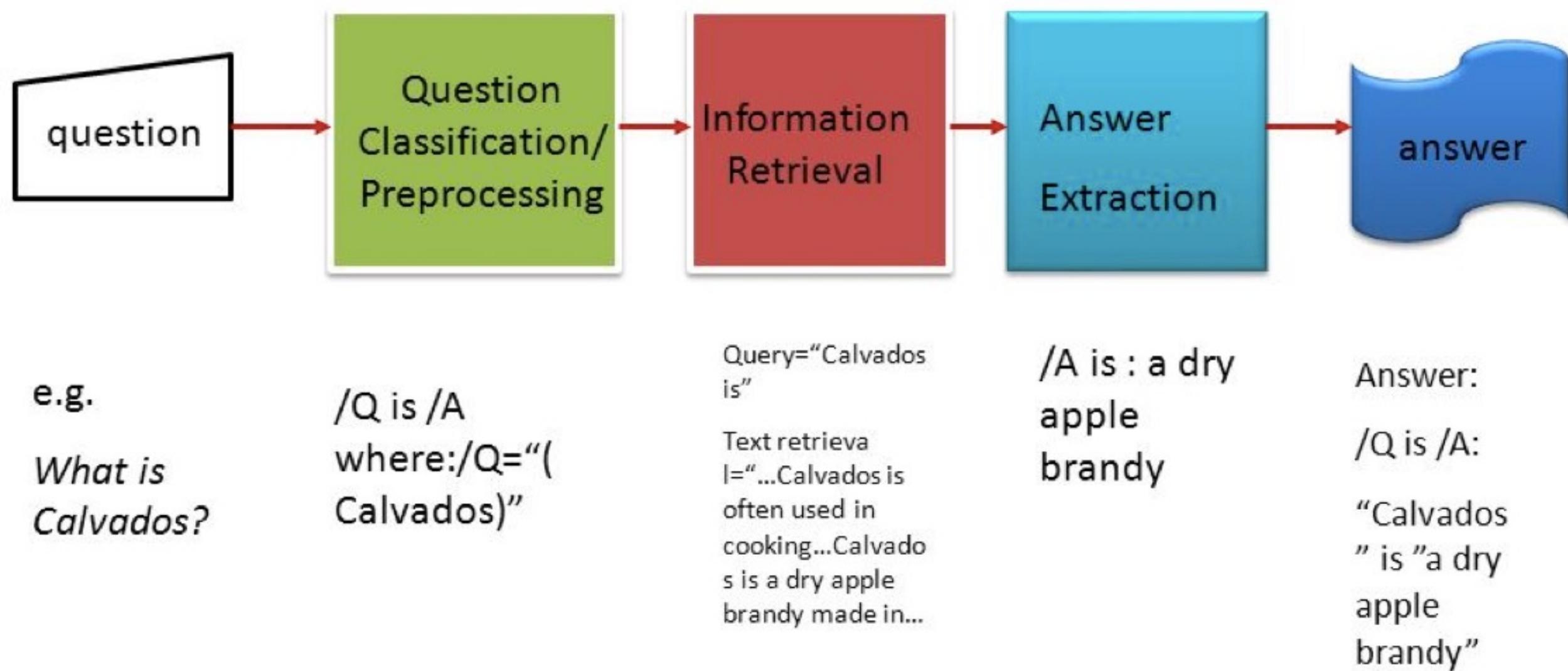
6. Автоматический поиск различий в документах



NLP

Задачи

7. Вопросно-ответные системы (QA Systems), или системы с естественно-языковым интерфейсом



Used in Apple's Siri, Wolfram Alpha, IBM Watson.

NLP

Задачи

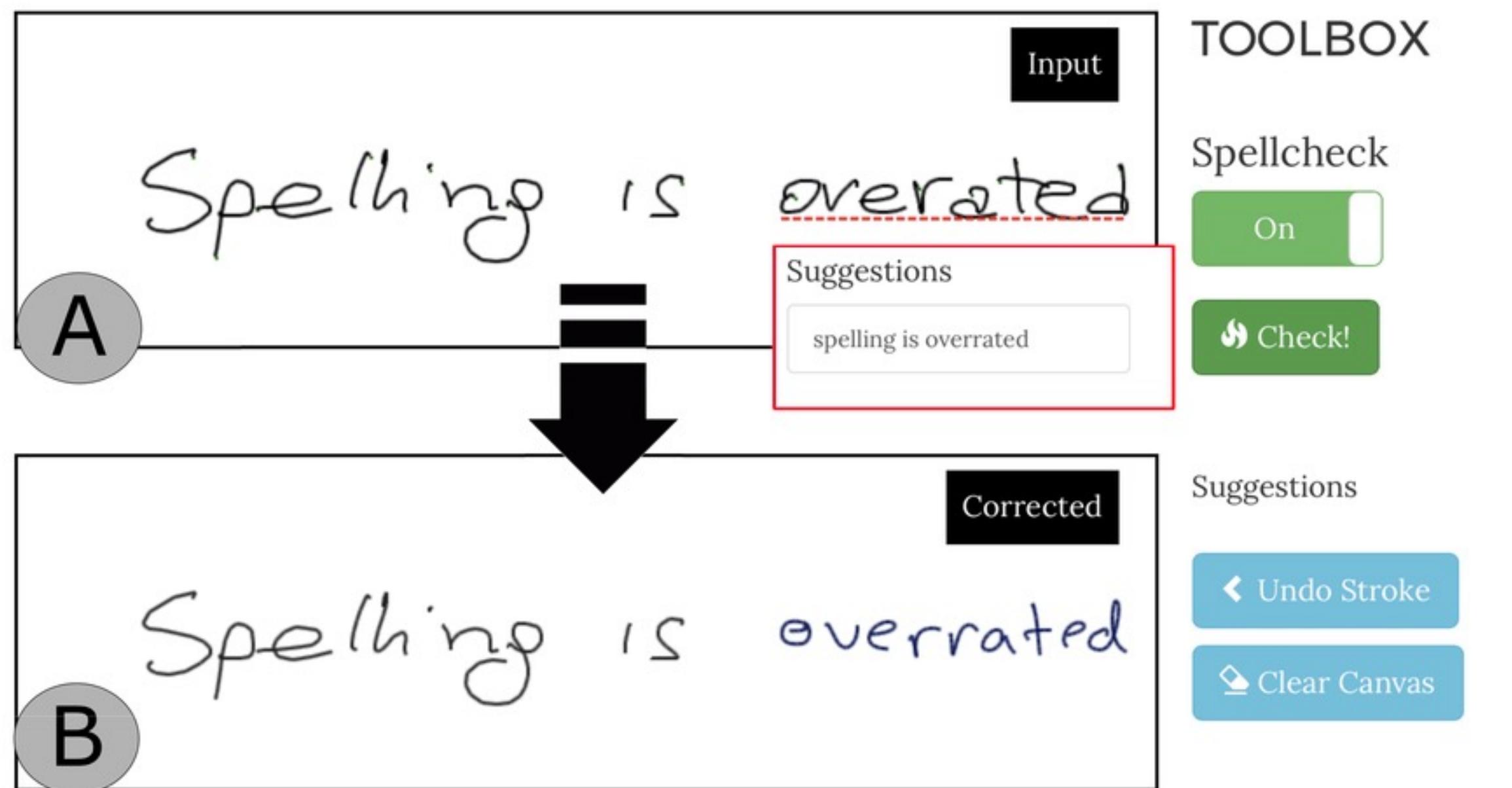
8. Диалоговые системы и чатботы



NLP

Задачи

9. Спелл-чекинг



NLP

Задачи

10. Генерация текста (try it!)

[Балабоба](#)



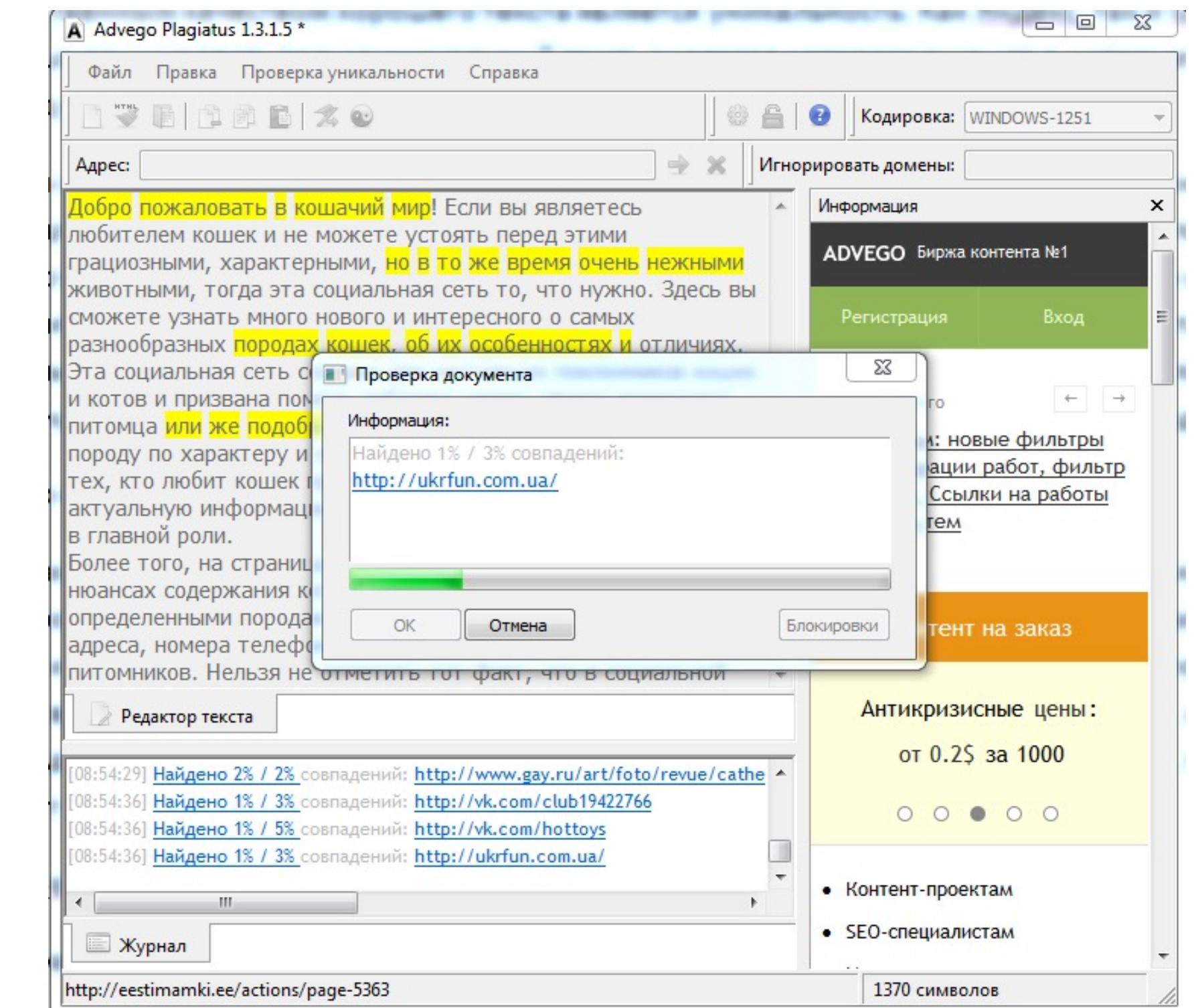
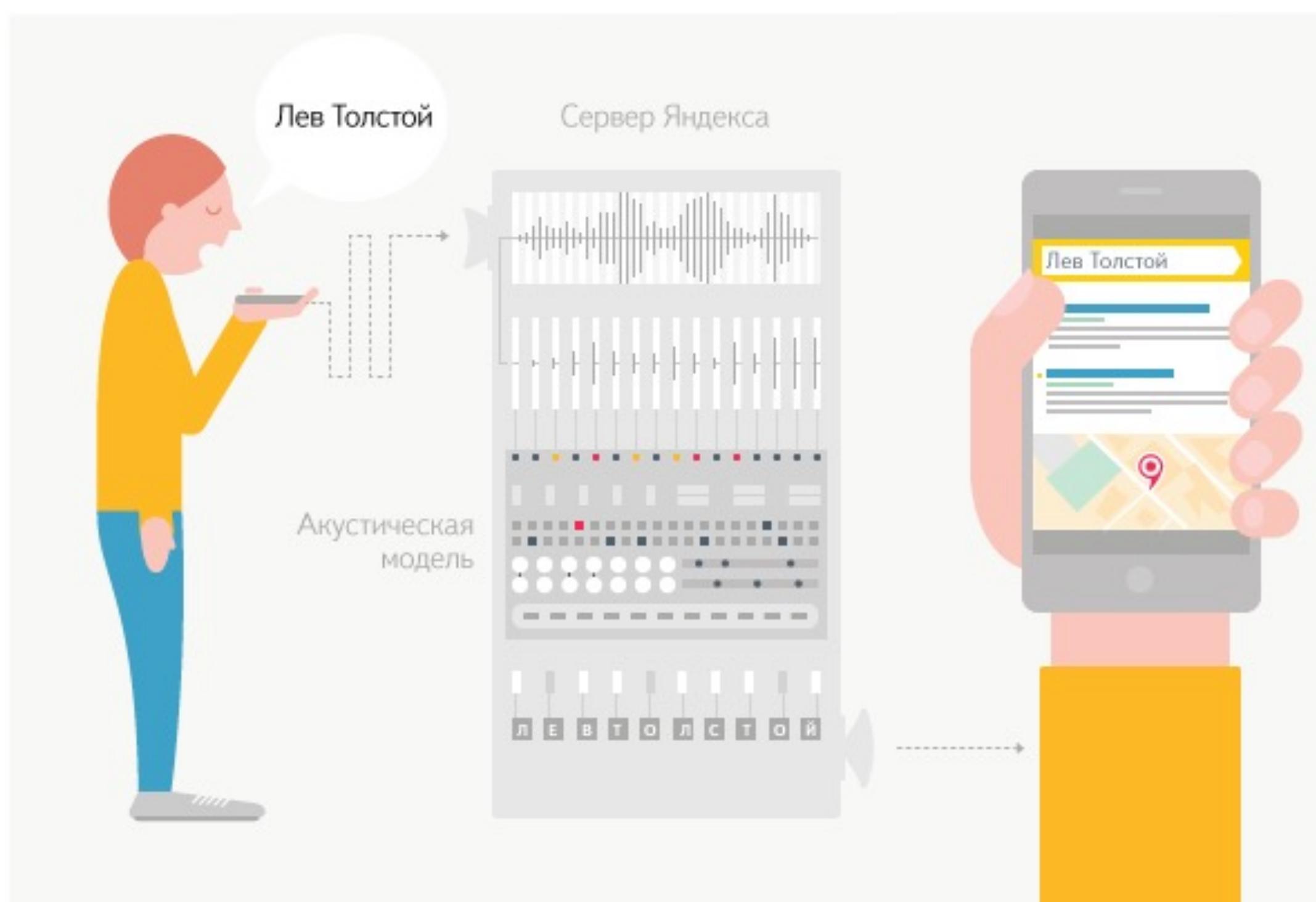
OpenAI GPT-3

NLP

Задачи

11. Автоматическая проверка текста на уникальность

12. Распознавание и синтез речи



NLP

Задачи

13. ...

Задача классификации пропаганды на SemEval

Генерация жестов по аудио голоса: <https://www.youtube.com/watch?v=xzTE5sobpFY>

В чем сложность?

В чем сложность?

Формулировки задач не очень сложные но сложный сам ЕЯ:

- Вариативность
- Омонимия
- Анафора
- Эллипсис
- Ирония, сарказм...

Методы работы с данными КЛ и NLP

- Правиловые (rule-based)
- Классическое машинное обучение (ML, МО, машинка)
- Глубокое обучение (нейронные сети, сетки, Deep learning, DL)
- Гибридные

Правиловый метод

Шаги

1. Определяем языковые правила, которые описывают поведение данных
2. Задаем порядок, как правила должны применяться к данным.
3. Пишем код (`if ... elif ... else ...`)

Проблемы?

Правиловый метод

Шаги

1. Определяем языковые правила, которые описывают поведение данных
2. Задаем порядок, как правила должны применяться к данным.
3. Пишем код (`if ... elif ... else ...`)

Проблемы?

- В ЕЯ много исключений, сложно формализовать все закономерности.
- Очень лингвоспецифично
- и задаче-специфично - плохо масштабируются

Правиловый метод

Задачи

1. Сегментация (деление на предложения)
2. Токенизация
- 3.*Лемматизация
- 4.*PoS-tagging (PoS-тэггинг, частеречная разметка)
- 5.*Синтаксический парсинг (см. СинТагРус)

Правиловый метод

Задачи

1. Сегментация (деление на предложения)
2. Токенизация (деление на слова токены)
- 3.*Лемматизация
- 4.*PoS-tagging (PoS-теггинг, частеречная разметка)
- 5.*Синтаксический парсинг (см. СинТагРус)

Вывод:

Используется для конкретных небольших задач в проекте.

Остальные более сложные задачи лучше решаются методами МО и глубокого обучения.

Классическое МО

Шаги

1. Размечаем данные
2. Подбираем признаки (что важно, а что нет?)
3. Обучаем модель
4. Тестируем на данных, не исп. для обучения

Подробнее о МО мы поговорим на другом занятии. Пока важна лишь концепция.

Классическое МО

Шаги

- 1. Размечаем данные**
- 2. Подбираем признаки (что важно, а что нет?)**
- 3. Обучаем модель**
- 4. Тестируем на данных, не исп. для обучения**

Проблемы?

Классическое МО

Шаги

1. Размечаем данные
2. Подбираем признаки
3. Обучаем модель
4. Тестируем на данных, не исп. для обучения

Проблемы?

Размеченные данные (много данных и качественная разметка!)

Данные меняются (напр. жанр текста) -> качество хуже

Выбрать признаки (что важно, а что нет?)

Нейронные сети

Шаги

- 1. Размечаем данные**
- 2. Обучаем модель**
- 3. Тестируем на данных, не исп. для обучения**

Проблемы?

Нейронные сети

Шаги

1. Размечаем данные

2. Обучаем модель (часто уже предобучена, т.е. не с нуля)

3. Тестируем на данных, не исп. для обучения

Проблемы?

Размеченные данные (много данных и качественная разметка!)

Данные меняются (напр. жанр текста) -> качество хуже

~~Выбрать признаки (что важно, а что нет?)~~

Признаки

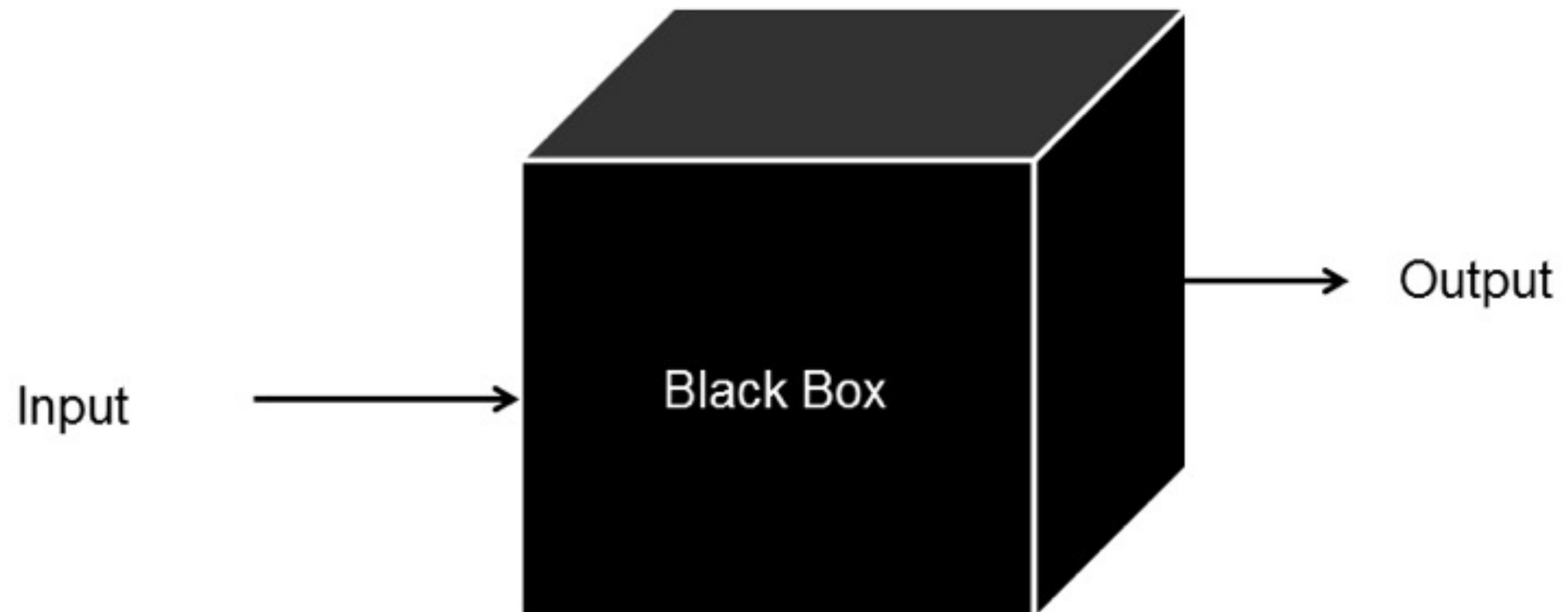
Или как это работает и что важно?

Мы не выбираем признаки



Проще, не ошибемся в подборе
признаков, не тратим
человеческий ресурс (feature
engineering), быстрее

Black box problem:
Как это работает?
Что происходит с объектом внутри
нейронной модели?
Что важно при решении этой
задачи, а что нет?



Internal behavior of the code is unknown

МО и нейронные сети

Кто кого?

- 2-4 года назад (в зависимости от задачи) классические методы МО были не хуже нейросетей. Сейчас нейросети практически для всех задач дают более высокие результаты.
- Классические методы для каждой задачи использовали разные архитектуры и наборы признаков
- Архитектуры нейросетей для разных задач различаются меньше, а признаки часто вообще не отличаются: сформировался пайплайн NLP

Гибридный подход

- МО / DL + правиловые методы (словари, знания о языке...)

Wordform	Right lemma	BERT	ELMo
потерь	потеря	потеть	потерья
подсел	подсесть	подйти	подсеть
льдах	лед	льер	льд
прилечу	прилететь	прилестить	прилечуть
пою	петь	повать	поть
берите	брать	беТЬ	берить
бегите	бежать	бяться	бегять
шипящими	шипеть	шипить	шипть
стань	стать	станть	стть
зажгли	зажечь	зжечь	зажгть

Table 2. Examples of hallucination errors

Нейронные сети

Пайплайн

- Сегментация

Нейронные сети

Пайплайн

- Сегментация
- Токенизация

Нейронные сети

Пайpline

- Сегментация
- Токенизация
- Лемматизация / стемминг

Нейронные сети

Пайpline

- Сегментация
- Токенизация
- Лемматизация / стемминг

Мы не можем подать на вход слова (буквы), можем только цифры

Нейронные сети

Пайpline

- Сегментация
- Токенизация
- Лемматизация / стемминг
- Слово -> вектор (эмбеддинг)

Эмбеддинги

Лирическое отступление

Index	Job
1	Police
2	Doctor
3	Student
4	Teacher
5	Driver



One hot encoded data
[1 0 0 0 0]
[0 1 0 0 0]
[0 0 1 0 0]
[0 0 0 1 0]
[0 0 0 0 1]

Нейронные сети

Пайpline

- Сегментация
- Токенизация
- Лемматизация / стемминг
- Слово -> вектор (эмбеддинг)

Нейронные сети

Пайpline

- Сегментация
- Токенизация
- Лемматизация / стемминг
- Слово - > эмбеддинг
- (PoS-tagging)

Нейронные сети

Пайплайн

- Сегментация
- Токенизация
- Лемматизация / стемминг
- Слово - > эмбеддинг
- (PoS-tagging)
- (Морфологические признаки, они же features)

Нейронные сети

Пайплайн

- Сегментация
- Токенизация
- Лемматизация / стемминг
- Слово -> эмбеддинг
- (PoS-tagging)
- (Морфологические признаки, они же features)
- (Синтаксический парсинг)