



Программы, приложения, web-сервисы

Глотова Т.В.

Пензенский государственный университет,
Пенза, Россия



ПРИКЛАДНЫЕ ЗАДАЧИ

- Машинный перевод
- Информационный поиск
- Классификация и кластеризация текстов
- Реферирование и аннотирование текстов
- Формирование ответов на вопросы
- Автоматизация подготовки и редактирование текстов
- Извлечение информации из текстов
- Генерация текстов на ЕЯ
- Организация диалога с пользователем на ЕЯ
- Обучение ЕЯ
- Определение тональности
- Распознавание и синтез звучащей речи



Прикладная лингвистика

обработка ЕЯ - междисциплинарная область:

- Компьютерная лингвистика
- Смежные области исследований
 - Лингвистика
 - Математика
 - Информатика (Computer Science)
 - Искусственный интеллект

ABBYY

<https://www.abbyy.com/ru-ru/>

ABBYY — российская компания-разработчик решений в области распознавания текстов и лингвистики.

Наиболее известные продукты — программа для распознавания текстов ABBYY FineReader, система потокового ввода данных ABBYY FlexiCapture и электронные словари ABBYY Lingvo.

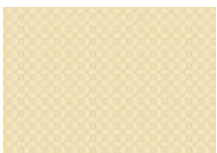
- «Только в российских СМИ в день выходит более 500 тысяч сообщений*, постоянно обновляются базы данных и информационно-поисковые системы. Каждую минуту пользователи отправляют более 50 тысяч ссылок в Facebook и публикуют 350 тысяч постов в Twitter».
- ABBYY является организатором крупнейшей конференции по компьютерной лингвистике в России — «Диалог», проводит открытые семинары по компьютерной лингвистике ABBYY Open.
- <https://www.abbyy.com/ru-ru/products/>



Этапы работы ABBYY Compeno



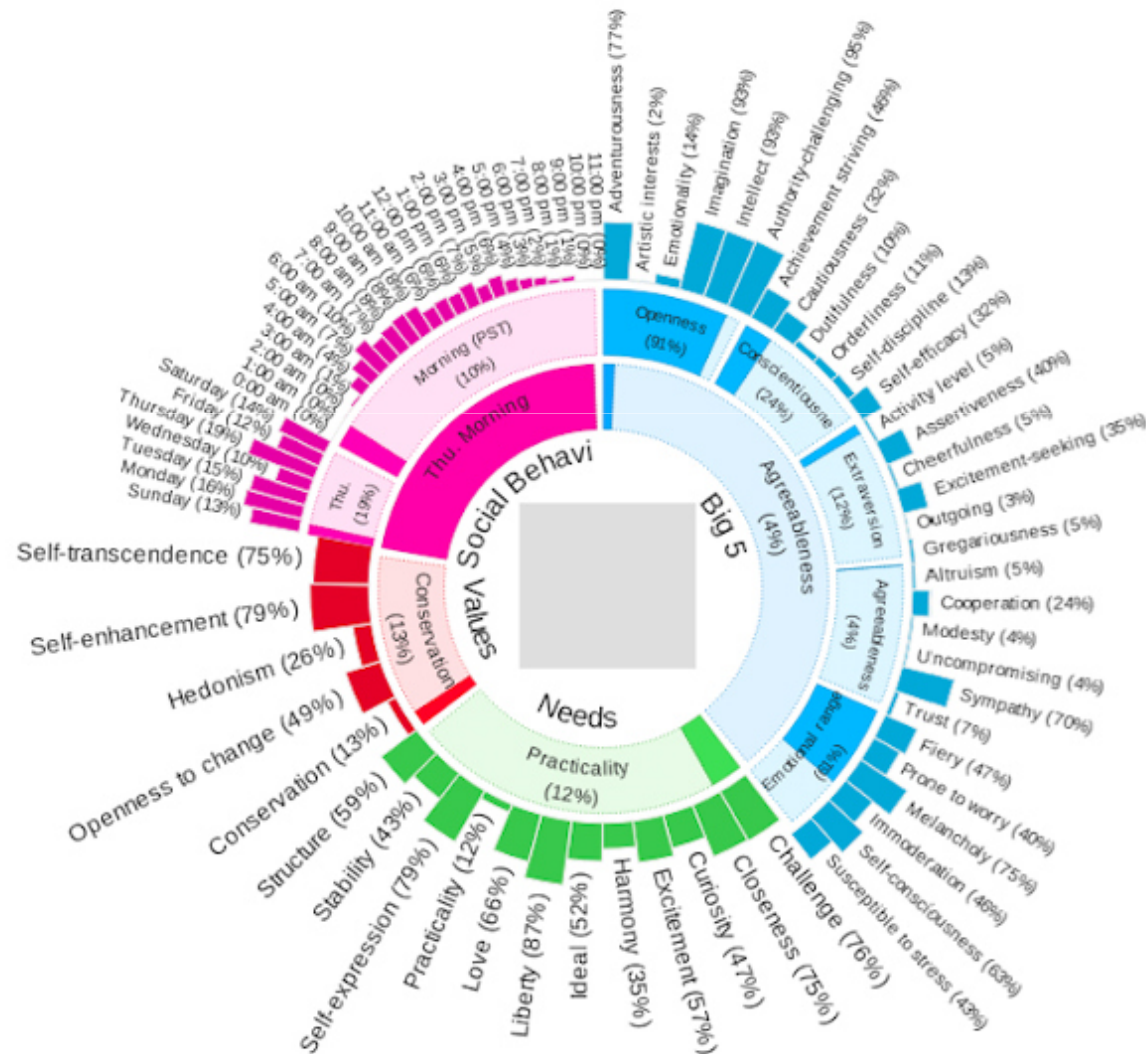
Расскажите о своей задаче, и мы предложим



IBM Watson

- облачная аналитическая система, работающая с гигантскими объемами собранной информации
- использует современные компьютерные технологии: машинное обучение, компьютерная лингвистика, онтологические построения и высокопроизводительные вычисления.
- Ключевые умения IBM Watson
 - Понимание естественного языка.
 - Построение гипотез на основе обработанных данных.
 - Обучение в процессе работы.
 - Вынесение рекомендации, сопровождая ее фактами, на которых основан вывод.

IBM Watson: Покажи мне свой аккаунт в социальных сетях, и я скажу тебе, кто ты





IBM

- Получите информацию, которая поможет вам быстрее принимать решения

<https://www.ibm.com/ru-ru/customer-engagement/digital-marketing/customer-insights>

- Выявление и использование практической терминологии с помощью IBM Watson Content Analytics

<https://www.ibm.com/developerworks/ru/library/ba-watson-dictionary/>

- Сбор и анализ данных из социальных сетей абсолютно без написания кода

<https://www.ibm.com/developerworks/ru/library/ba-collect-analyze-social-data-app/>



Яндекс

- Помимо поисковых услуг, Яндекс предлагает ряд служб и инструментов для различных нужд пользователей.
- Большинство из них — веб-приложения, требующие от пользователя только наличия браузера, в котором они работают, и интернет-подключения. Это позволяет использовать данные в любой точке планеты и не быть привязанным к одному компьютеру.
- Целый ряд служб — [Яндекс.Поиск](#), [Яндекс.Карты](#), [Яндекс.Маркет](#), [Поиск по блогам](#), [Народ.ру](#), [Яндекс.Пробки](#) — лидируют на российском рынке
- Помощник Алиса

Элементы Яндекса

Свободно распространяемое программное обеспечение в виде надстройки для браузеров от компании «Яндекс», позволяющее существенно расширить их функциональность и повысить удобство использования.

«Элементы» представляют собой ноу-хау, - они, в отличие от традиционных панелей инструментов, могут появляться по мере необходимости и не перегружают интерфейс.

<https://yandex.ru/soft/element/>

https://translate.yandex.ru/?utm_source=wizard

<https://yandex.ru/soft/element/translate/>

Наряду с уже знакомой функцией перевода отдельных слов, Элементы предложат вам перевести страницу целиком, если она написана на иностранном языке:

Сервис подбора слов

- Подбор слов (wordstat) — это сервис, который помогает получить информацию о запросах пользователей Яндекса. Например, он позволяет узнать, сколько людей в месяц ищут ту или иную фразу, и посмотреть запросы, похожие по смыслу на вашу фразу.
- <https://yandex.ru/support/direct/keywords/wordstat.html>
- <https://wordstat.yandex.ru/>



Wordstat

- цифры, которые располагаются рядом с каждым запросом, представляют собой предварительную оценку количества месячных показов.
- Сервис имеет специальные операторы, позволяющие проверить ключевые слова в Яндексе. Например, «кавычки» позволят учитывать показы исключительно по слову и словоформам без дополнительных слов.
- При запросе «семантическое ядро» учитывались показы по различным формам фразы, как, например «семантическому ядру», «семантическим ядром», но не «семантическое ядро веб-ресурса». Восклицательный знак позволяет учитывать слово именно в таком варианте написания.

Яндекс метрика

- <https://yandex.ru/blog/metrika>
Яндекс.Метрика, сервис, разработанный специально для администраторов сайтов и веб-мастеров. Данный сервис позволяет анализировать статистику посещаемости, оценивать действия посетителей.
- Помимо всего прочего в рамках метрики реализована функция мониторинга, которая позволяет узнать об имеющихся проблемах доступности вашего сайта с помощью SMS.

Сайты и ресурсы

- Компьютерные ресурсы школы лингвистики

<https://ling.hse.ru/resources>

- NLPub — каталог ресурсов для обработки естественного языка <https://nlpub.ru/>

<https://nlpub.ru/%D0%9E%D0%BI%D1%80%D0%B0%D0%BI%D0%BE%D1%82%D0%BA%D0%B0%D1%82%D0%B5%D0%BA%D1%81%D1%82%D0%B0>

- Информационные технологии: электронное обучение, компьютерная филология Елисеева О.Е. Белорусский ГУ

<http://it.lang-study.com/veb-sajty-filologicheskoy-i-lingvisticheskoy-tematiki/>

- Корпусная лингвистика СПбГУ

<http://corpora.iling.spb.ru/>

Каталог лингвистических программ и ресурсов в Сети

<https://rvb.ru/soft/catalogue/>

<https://rvb.ru/soft/catalogue/c01.html>

Инструменты

- **Gephi**: инструмент для обработки данных из домена G Suite

<http://edublogru.blogspot.com/2017/03/gephi-g-suite.html>

- **Gromoteur** - инструмент для лингвистов, который обеспечивает легкий доступ к текстовым корпусам. Он позволяет получать страницы из Интернета или из локальных файлов, обрабатывать их, анализировать и выводить результаты.

<http://gromoteur.ilpga.fr/#howtos>

- [gr\Instant Corpora_ Начало работы с Gromoteur.html](#)

Национальный корпус русского языка <http://ruscorpora.ru/>



НАЦИОНАЛЬНЫЙ КОРПУС
РУССКОГО
ЯЗЫКА

главная

архив новостей

поиск в корпусе

что такое корпус?

состав и структура

статистика

графики

частоты

морфология

обороты

синтаксис

семантика

параметры текстов

studiorum

forum

Национальный корпус русского языка

На этом сайте помещен корпус современного русского языка общим русского языка — это информационно-справочная система, основ электронной форме.

Корпус предназначен для всех, кто интересуется самыми разными во профессиональных лингвистов, преподавателей языка, школьников русский язык.

Развитие подкорпусов НКРЯ (основного, поэтического, параллельного, году осуществлялось при поддержке РГНФ, проекты № 15-04-12018 « НКРЯ» и № 14-04-12012 «Корпус диалектных текстов Национального разметка».

[Как пользоваться Корпусом \(инструкция в формате PDF\)](#)

[Подробнее о корпусе](#)

Новости проекта

3 апреля 2018 года

Объём [латышско-русского и русско-латышского](#) параллельного корпуса словоупотреблений. Объём [бурятско-русского и русско-бурятского](#) параллельного корпуса словоупотреблений.

Natural Language Toolkit (NLTK)

NLTK <http://nltk.org/>

- пакет библиотек и программ для символьной и статистической обработки естественного языка, написанных на языке программирования [Python](#). Сопровождается обширной документацией, включая книгу с объяснением основных концепций, стоящих за теми задачами обработки естественного языка, которые можно выполнять с помощью данного пакета.
- NLTK хорошо подходит для изучающих компьютерную лингвистику или близкие предметы, такие как эмпирическая лингвистика, когнитивистика, искусственный интеллект, информационный поиск и машинное обучение.
- NLTK является свободным программным обеспечением.

Freeling

<http://nlp.isi.upc.edu/freeling>

- Freeling это библиотека для обработки текстов. Написана на C++, за счет чего имеет хорошую скорость работы. А так же имеет API почти на всех языках.
- Основные возможности FreeLing:
 - • Токенизация
 - • Морфологический анализ
 - • Приведение слова к начальной форме
 - • Определение части речи, в том числе — неизвестного слова
 - • Нормализация и определение дат и чисел
 - • Распознавание и классификация сущностей



Инструменты проверки английского языка

- Spell checker: check grammar and spelling for English texts

<http://www.reverso.net/spell-checker/english-spelling-grammar/>

- <https://www.multitran.ru/>
- <https://www.grammarly.com/>



Компьютерная грамматика русского языка: лексика, морфология, синтаксис

<http://solarix.ru>

- **Грамматический Словарь Русского Языка**
- **Парсер - лемматизатор, морфологический и синтаксический анализатор текста**
- **SDK Грамматического Словаря Русского Языка**
- **SQL Словарь Русского Языка**



Eureka Engine

- Модуль автоматического определения языка (LanguageDetector)
- Модуль автоматического определения тональности (SentiFinder)
- Модуль автоматической классификации текстов (TextClassifier)
- Модуль автоматического определения именованных сущностей (NER)
- Модуль нормализации слов русского языка
- Модуль морфологического анализа русского языка
- <http://eurekaengine.ru/ru/demo/>



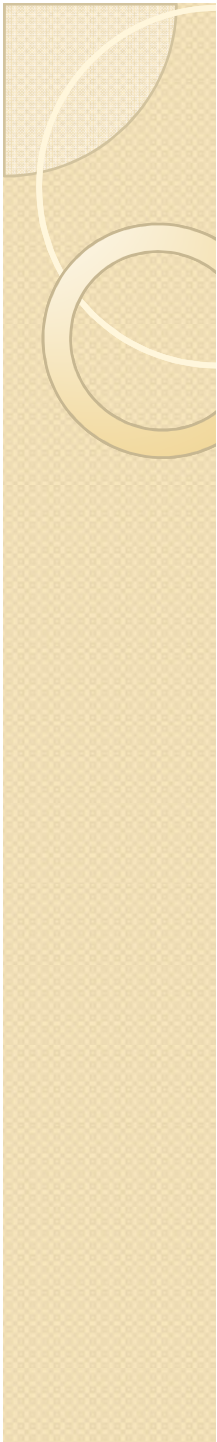
Морфологический анализатор

- **Py morphology**

<https://pythonhosted.org/pymorphy/intro.html>

- Морфологический анализатор
pymorphy2

<https://pymorphy2.readthedocs.io/en/latest/index.htm>



Морфологический анализатор rutmorphy2

- приводит слово к нормальной форме (например, “люди -> человек”, или “гулял -> гулять”).
- ставит слово в нужную форму. Например, ставить слово во множественное число, менять падеж слова и т.д.
- возвращает грамматическую информацию о слове (число, род, падеж, часть речи и т.д.)
- При работе используется словарь [OpenCorpora](#); для незнакомых слов строятся гипотезы. Скорость работы - от нескольких тыс слов/сек до > 100тыс слов/сек (в зависимости от выполняемой операции, интерпретатора и установленных пакетов); потребление памяти - 10...20Мб; полностью поддерживается буква ё.

Сравнение текстов

<http://www.topwriter.ru/comparison/%20http://q99.it/8oxsR5p/>

TopWriter β — сам себе рерайтер!

Пример работы рерайтера

Рабочее окно состоит двух полей — «Было» и «Стало». В поле «Было» вносится исходный текст, подлежащий преобразованию.

Было:

устал. Слуга подождал четверть часа, опять стал будить, опять не добудился. Стал советоваться с другими слугами, с буфетчиком. "Уж не случилось ли с ним чего?" - "Надо выломать двери". - "Нет, так не пойдет: дверь ломать надо с полицией". Решили попытаться будить еще раз, посильнее; если и тут не проснется, послать за полицией. Сделали последнюю пробу; не добились; послали за полицией и теперь ждут, что увидят с нею.

Часам к 10 утра пришел полицейский чиновник, постучался сам, велел слугам постучаться, - успех тот же, как и прежде. "Нечего делать, ломай дверь, ребята".

Дверь выломали. Комната пуста. "Загляните-ка под кровать" - и под кроватью нет проезжего. Полицейский чиновник подошел к столу, - на столе лежал лист бумаги, а на нем крупными буквами было написано:

"Ухожу в 11 часов вечера и не возвращусь. Меня услышат на Литейном мосту

Стало:

прибыл господин с чемоданом, взял номер, для прописки отдал свой паспорт, потребовал себе котлетку и чаю, сказал, чтоб его не тревожили вечером, потому что он собирается вздремнуть и устал, однако чтобы завтра непременно разбудили в 8 часов, потому что у него есть неотложные дела, закрыл дверь номера и, пошумев ножом и вилкою, пошумев чайным прибором, мгновенно стих, - видно, заснул. Наступило утро; в 8 часов человек постучался к вчерашнему приезжему - приезжий не полагает голоса: слуга



Главред <https://glvrd.ru/>

- [Главред](https://glvrd.ru/) — это сервис для улучшения текста.
- «Главред помогает находить в тексте стоп-слова: языковой мусор, рекламные и журналистские штампы, признаки плохого синтаксиса и канцелярит. Без этих слов текст становится понятным, коротким и информативным. Мы называем это информационным стилем».



Проект Tatoeba

<https://tatoeba.org/rus>

- Проект Tatoeba — это обширная база данных предложений и их переводов, постоянно пополняющаяся усилиями тысяч добровольных участников.
- Проект Tatoeba дает инструмент, с помощью которого можно увидеть как то или иное слово используется в контексте предложения.
- «Вы задаете интересующие вас слова и получаете результат, содержащий их и их переводы на желаемые языки»
- Слово Tatoeba (по-японски *например*) отражает эту концепцию.

Ассоциативный поиск

- <https://wordassociations.net/ru>
- Проект «Сеть словесных ассоциаций» дарит вам возможность просмотра ассоциаций, возникающих у человека, с определённым словом.
- Поиск словосочетаний в Интернете <https://visualworld.ru/>



Praat

<http://www.fon.hum.uva.nl/praat>

- Основные функции : сегментация потока речи на примере предложения на синтагмы, слова и звуки.
- Данная программа позволяет получить осциллограмму желаемой аудиозаписи, а также графические и числовые данные о частоте основного тона, интенсивности, формантах, пульсах речи, а также ее спектр.
- Программа PRAAT используется для фонетической обработки материала – заполнения аннотаций на фонетических уровнях.
- Программа используется как в научных исследованиях, так и в обучении

ELAN

- программа лингвистического аннотирования ELAN является крайне удобной для представления многоуровневой разметки разнородного материала аудио и видео, хранит данные в удобном для их последующей автоматической обработки виде.
- <https://tla.mpi.nl/tools/tla-tools/elan/screenshots/>

Литература

1. IBM Watson: Покажи мне свой аккаунт в социальных сетях, и я скажу тебе, кто ты <https://habr.com/company/ibm/blog/388891/>
2. <http://elib.spbstu.ru/dl/2/3250.pdf/download/3250.pdf>
3. МГУ ВМК Большакова Компьютерная лингвистика: методы, ресурсы, приложения Е.И. http://www.docme.ru/doc/359479/komp._yuternaya-lingvistika--metody--resursy—prilozheniya
4. <https://habr.com/company/yandex/blog/146618/>
5. NLTK по-русски <https://ailev.livejournal.com/1006416.html>
6. Pymorphy <https://pythonhosted.org/pymorphy/intro.html>
7. Запрос информации по лингвистическому анализу <https://groups.google.com/forum/#!topic/pymorphy/GetdEELx3il>
8. Морфологический анализатор pymorphy2 <https://pymorphy2.readthedocs.io/en/latest/index.html>
9. Morphological Analyzer and Generator for Russian and Ukrainian Languages статья <https://arxiv.org/pdf/1503.07283v1.pdf>
10. Обработка естественного языка. Полезные инструменты <http://www.pvsm.ru/python/13305/print/>
11. Корпусная лингвистика Бочаров В. http://opencorpora.org/doc/presentations/2011_Witology.pdf