

Графематический и морфологический анализ текстов

Графематический анализ (токенизация)

- Разделение текста на слова, разделители
- Выделение устойчивых оборотов, не имеющих словоизменительных вариантов
- Выделение предложений
- Выделение абзацев
- Выделение дат
- Определение языка слова (русский, нерусский)
- Определение формата написания слова (прописные, строчные буквы)

Сегментация текста на слова

Англ. Tokenization

Принципиальные возможности

- в орфографии данного языка предусмотрены пробелы между словами;
- в орфографии данного языка нет пробелов или иных разделителей между словами.

Сегментация на слова текста с пробелами

Осложняющие факторы:

- сегменты текста между пробелами требуют переразложения
 - буду (часто) писать; железная дорога; с разбегу;
 - *Du holst mich ab* (нем.)
- словоформы могут разделяться не только пробелами
 - *наконец-то* (vs *кто-то, во-первых, по-моему*),
 - *they're* и *isn't* (vs *friend's*), *and/or* (vs *accept/reject*)

Сегментация текста без пробелов между словами

[トップ](#) [主要](#) [経済](#) [企業](#) [株・為替](#) [国際](#) [政治](#) [社会](#) [スポーツ](#) [新製品](#) [リ](#)

ソニー、金融子会社10月上場・3000億円調達、今年最大に

ソニーの全額出資の金融子会社ソニーフィナンシャルホールディングス（SFH）の上場日程が固まった。東京証券取引所第一部に10月上場し、公募・売り出しを合わせた株式の公開規模は3000億円前後で今年最大の上場案件になる。ソニーは売却で得た資金を主力のエレクトロニクス（電機）部門の強化に充て、選択と集中を加速する。

今週半ばにも発表する。上場は10月上旬を予定。ソニーは保有している株式のうち3割強を売り出すほか、SFHが新株を発行する。2006年11月上場したあおぞら銀行（約3800億円）以来の大型案件で、上場時の時価総額は1兆円前後に達するとの

日経ブ
Bro

シ
の

映像二

Разбиение текста на предложения

- Приложения
 - синтаксический анализ (парсеры)
 - системы автоматического реферирования
 - машинный перевод
 - Извлечение терминов...
 - Информационный поиск: оператор поиска слов внутри предложения

Результат: текст, разбитый на предложения

«Наивная» сегментация

По знакам препинания:

- точка,
- восклицательный и вопросительный знаки

«Наивная» сегментация

По знакам препинания:

- точка,
- восклицательный и вопросительный знаки

В связи с этим первый интервал пробегов был принят равным 350...700 тыс. км (середина интервала - 525 тыс. км), второй интервал -- 700...1050 тыс. км (середина интервала - 875 тыс. км) и третий интервал 1050...1400 тыс. км (середина интервала -- 1225 тыс. км).

«Наивная» сегментация

По знакам препинания:

- точка,
- восклицательный и вопросительный знаки

Политический кризис в Сирии представляет опасность для ближневосточного региона и всего мира. Такое мнение высказал спецпосланник ООН и Лиги арабских государств (ЛАГ) **Л. Брахими**, передает Reuters.

Практические решения

- Предложение должно содержать буквы
- Предложение должно начинаться с заглавной буквы
- Сокращения (из списка) требуют «особого внимания»
 - г., тыс., млн., ул ...
- Отдельные большие буквы: А.Б. Иванов
- ...

Операции со словами

- Словоизменение
 - изменение одного и того же слова
 - в основном есть регулярные правила в языке
 - Лесной, лесная, лесного...
- Словообразование
 - образование новых слов
 - Лес, лесник, лесничество...
- В информационном поиске очень важно уметь определять формы одного и того же слова (словоизменение)

Морфологический анализ

- Морфологический анализ текста осуществляет приведение словоформ, встречающихся в тексте, к нормальному (словарному) виду и определяет морфологические характеристики словоформы
- Нормальная форма (=словарная форма=лемма):
 - Существительные, прилагательные
 - им. падеж
 - ед. число
 - мужской род
 - Глагол: инфинитив
- Упрощенная процедура: лемматизация (=восстановление нормальной формы)
- Обратная процедура: морфологический синтез

Морфологические характеристики словоформ русского языка

- Имя существительное:
6 падежей * 2 числа
- Имя прилагательное:
6 падежей * 2 числа (в ед.ч. 3 рода)
+ 4 краткие формы
+ степени сравнения
- Глагол:
(неопр.ф. + личные формы изъяв.накл. +
повел.накл. + прич. + деепр.) * 2 вида
- Неизменяемые части речи...

Обработка словоформы: морфологический анализ

исследовать	{исследовать} + +Неопр.ф.
исследую	{исследовать} + + Наст., Буд. вр. + Ед.ч. + 1 л.
исследуешь	{исследовать} + + Наст., Буд. вр. + Ед.ч. + 2 л.
исследует	{исследовать} + + Наст., Буд. вр. + Ед.ч. + 3 л.
...	
исследовал	{исследовать} + + Прош. вр. + Ед.ч. + М р.
исследовала	{исследовать} + + Прош. вр. + Ед.ч. + Ж р.
...	

Порождение словоформы: морфологический синтез

{исследовать} + Неопр.ф.	исследовать
{исследовать} + Наст. вр. + Ед.ч. + 1 л.	исследую
{исследовать} + Наст. вр. + Ед.ч. + 2 л.	исследуешь
{исследовать} + Наст. вр. + Ед.ч. + 3 л.	исследует
...	
{исследовать} + Буд. вр. + Ед.ч. + 1 л.	исследую, буду исследовать
{исследовать} + Буд. вр. + Ед.ч. + 2 л.	исследуешь , будешь исследовать
{исследовать} + Буд. вр. + Ед.ч. + 3 л.	исследует, будет исследовать
...	
{исследовать} + Прош. вр. + Ед.ч. + М р.	исследовал
{исследовать} + Прош. вр. + Ед.ч. + Ж р.	исследовала
...	

Морфологический анализ и лемматизация

исследовать	{исследовать} + +Неопр.ф.
исследую	{исследовать} + + Наст., Буд. вр. + Ед.ч. + 1 л.
исследуешь	{исследовать} + + Наст., Буд. вр. + Ед.ч. + 2 л.
исследует	{исследовать} + + Наст., Буд. вр. + Ед.ч. + 3 л.
...	
исследовал	{исследовать} + + Прош. вр. + Ед.ч. + М р.
исследовала	{исследовать} + + Прош. вр. + Ед.ч. + Ж р.
...	

исследовать	{исследовать}
исследую	{исследовать}
исследуешь	{исследовать}
исследует	{исследовать}
...	
исследовал	{исследовать}
исследовала	{исследовать}
...	

Точнее: типы морфологического анализа

- Лемматизация – приведение к нормальной форме
 - Лесной, лесного, лесному->лесной
 - леса -> лес
 - Танцующая -> танцевать
- Стемминг –выделение псевдоосновы
 - Лесной, лесного, лес, лесистый -> лес
 - Система, системный, систематизировать->
 - систем

Методы морфологического анализа

а) словарный

- со словарем словоформ
 - Каждой словоформе поставлена в соответствие основа или лемма
- со словарем основ

б) бессловарный (фактически – со словарем псевдоокончаний)

+ анализ по аналогии («предсказание»)

Стемминг:
Алгоритм Портера Snowball

- <http://snowball.tartarus.org/algorithms/russian/stemmer.html>
- Список служебных слов
 - Союзы, предлоги, наречия, частицы
- Для остальных слов отделяются псевдоокончания
- Стеммеры были созданы для распространенных индоевропейских языков

Анализ алгоритмом Портера

- *Противоестественном*
- Рассматривается фрагмент слова после гласной (RV) :
тивоеестественном
 - В RV должна остаться хотя бы одна гласная
 - К фрагменту RV применяются заданные списки окончаний
- При нескольких вариантах выбирается наиболее длинное окончание

Окончание прилагательных в алгоритме Портера

- ее (**ee**) ие (**ie**) ые (**ye**) ое (**oe**) ими (**imī**) ыми (**ymī**) ей (**eī**) ий (**iī**) ый (**yī**) ой (**oī**) ем (**em**) им (**im**) ым (**ym**) ом (**om**) его (**ego**) ого (**ogo**) ему (**emu**) ому (**omu**) их (**ikh**) ых (**ykh**) ую (**uiu**) юю (**iuiu**) ая (**aia**) яя (**iaia**) ою (**oiu**) ею (**eiu**)

Словарные морфологии

Словарь Зализняка

1977

- «Грамматический словарь русского языка»

Автор – Андрей Анатольевич Зализняк (с 1997 г. академик РАН)

100тыс. входов - основа большинства компьютерных морфологий РЯ:

Проблемы:

- Автомобилестроения – мн.ч.
- Финансов – кр. Форма для финансовый
- При – пря
- Много старых слов
- Отсутствуют новые слова

Фрагмент страницы словаря А.А.Зализняка

ТЕЧЬ

ж (жо): 8а, 8е, 8f'' — 47 | св (нсв): 8 — 118

утечь	св нл 8b/b (-к-), ё 0II
вытечь	св нл 8а (-к-) 0II
дичь	ж 8а
на́взничь	н
опричь	предл.
стричь	нсв 8b (-г-)
застричь	св 8b (-г-) 0II
настричь	св 8b (-г-) 0II
обстричь	св 8b (-г-) 0II
подстричь	св 8b (-г-) 0II
перестричь	св 8b (-г-) 0II
остричь	св 8b (-г-) 0II
достричь	св 8b (-г-) 0II
постричь	св 8b (-г-) 0II
простричь	св 8b (-г-) 0II
состричь	св 8b (-г-) 0II
расстричь	св 8b (-г-) 0II
отстричь	св 8b (-г-) 0II
выстричь	св 8а (-г-) 0II
застичь	см. засти́гнуть
настичь	см. насти́гнуть
пристичь	см. присти́гнуть
достичь	см. дости́гнуть
постичь	см. пости́гнуть
жёлчь	ж 8а [// желчь =]

сво́лочь	жо 8е
сволочь	св 8b/b (-к-) [// простореч. сволочить] 0I(-а-)
отволочь	св 8b/b (-к-) [// простореч. отволочить] 0I(-а-)
уволочь	св 8b/b (-к-) [// простореч. уволочить] 0I(-а-)
выволочь	св 8а (-к-) [// простореч. выволочить] 0I(-а-)
толочь	нсв 8b/b (-к-) Δ наст. тол- кú, толчёт, толкúт; <i>прош.</i> толók, толклá, толókший; <i>прич. страд.</i> толчённый
затолочь	св 8b/b (-к-) Δ <i>буд.</i> зато- л кú, -чёт, -кúт; <i>прош.</i> -ók, -клá, -ókший; <i>прич. страд.</i> -чённый
натолочь	св, <i>спряж. см.</i> затолочь
втолочь	св, <i>спряж. см.</i> затолочь
подтолочь	св, <i>спряж. см.</i> затолочь
перетолочь	св, <i>спряж. см.</i> затолочь
потолочь	св, <i>спряж. см.</i> затолочь
протолочь	св, <i>спряж. см.</i> затолочь
столочь	св, <i>спряж. см.</i> затолочь
растолочь	св, <i>спряж. см.</i> затолочь

то́чь-в-то́чь	н
запрячь	св 8b/b (-г-) 0II
перезапрячь	св 8b/b (-г-) 0II
напрячь	св 8b/b (-г-) 0II
поднапрячь	св 8b/b (-г-)
перенапрячь	св 8b/b (-г-) 0II
впрячь	св 8b/b (-г-) 0II
подпрячь	св 8b/b (-г-) 0II
перепрячь	св 8b/b (-г-) 0II
припрячь	св 8b/b (-г-) 0II
сопрячь	св 8b/b (-г-) 0II
спрячь	св 8b/b (-г-) 0II
распрячь	св 8b/b (-г-) 0II
отпрячь	св 8b/b (-г-) 0II
упрячь	св 8b/b (-г-) 0II
выпрячь	св 8а (-г-) 0II
наотмашь	н
ро́спашь	ж 8а
гуа́шь	ж 8а
плешь	ж 8а
флешь	ж 8а
брешь	ж 8а
ишь	част.; межд.
бишь	част.
вишь	част.
вишь	(насто без удара)

Схема морфологического анализа со словарем

- Для неслужебных слов:
- Выделить возможные окончания слова длиной от 0 до 3 СИМВОЛОВ
- Для каждого полученного окончания определить код окончания по таблице окончаний и номер флективного класса по словарю основ (лемм)
- Если номер флективного класса и номер окончания найдены, то проверить их согласованность по морфологической таблице
- Если согласованность подтверждается, то сохранить данный вариант

Процедура определения типовой парадигмы

- если слово оканчивается на *щийся*, то ТП 5;
- если слово оканчивается на *ин*, *ын*, то ТП 20;
- если слово оканчивается на *ов*, *ёв*, *ев*, то ТП 21;
- если слово оканчивается на *цый*, то ТП 6;
- если слово оканчивается на *ый*, то ТП 1;
- если слово оканчивается на *кий*, *гий*, *хий*, то ТП 3;
- если слово оканчивается на *ций*, то ТП 4;
- если слово оканчивается на *жий*, *ший*, *чий*, то ТП 4 или ТП 24;
- если слово оканчивается на *ий*, то ТП 2 или ТП 24;
- если слово оканчивается на *кой*, *гой*, *хой*, *жой*, *шой*, *чой*, *щой*, то ТП 8;
- если слово оканчивается на *ой*, то ТП 7.

Морфологический анализ на базе словаря: проблемы

- Дают максимально полный анализ словоформы
- На реальных текстах дают сбои (опечатки, уникальные слова)
- Не существует абсолютно полных словарей – лексика языка непрерывно пополняется
- Для примера – невозможно включить в словарь всю существующую терминологию, имена, фамилии и т.д.

Морфологический анализ слов, отсутствующих в словаре

Предсказание в морфологическом анализе

- Функциональное назначение предсказания – морфологический анализ слов (словоформ), отсутствующих в словаре
- Метод предсказания – выявление аналогий со словоформами, распознаваемыми имеющимся словарем

Алгоритм предсказания для новых слов

- 1) предсказание префиксального образования
- 2) предсказание по концовке, взятой из известных словоформ

Предсказание по префиксу

- Попытка найти существующую словоформу языка, которая максимально совпадала бы справа со входным словом.
- Если левая часть (потенциальный префикс) не длиннее М символов (пяти), а правая часть (совпавшая с известной словоформой) не короче N символов (четырех), то слово разбирается по образцу известной словоформы
– [евро]технологию, [супер]коньками
- Можно иметь список наиболее частых префиксов

Предсказание по концовке известной словоформы

Отделяются инвертированные концовки известных словоформа – длины К (пять букв),

Сопоставляются с морфологическими характеристиками:

- «анием» - как «ср. род, ед. ч., тв. пад.»

Такая строка заносится в исходный лексикон, если она встречается:

- не менее L раз (трех) и
- чаще конкурентов в пределах одной части речи

ВСЕГДА предусматривается разбор именем существительным, хотя бы неизменяемым.

Проблема морфологической омонимии

Пример:

На завод привезли **стекло**.

Масло **стекло** на пол.

Нес медведь, шагая к **рынку**,

На продажу меду **крынку**.

Вдруг на мишку - вот **напасть**!

Осы вздумали **напасть**.

Мишка с армией **осиной**

Дрался вырванной **осиной**.

Мог ли в ярость он не **впасть**,

Если осы лезли в **пасть**,

Жалили куда **попало**,

Им за это и **попало**.

Как решить?

Виды морфологической неоднозначности (=омонимии)

- Неоднозначность по леммам: косой, стали
 - Самая важная для информационного поиска
- Неоднозначность по частям речи: стали
- Неоднозначность по грамматическим характеристикам (падеж, число и др.)
 - Например, очень часто неоднозначность именительный vs. винительный падеж

Постморфологический анализ

- =предсинтаксический анализ
- Предназначен для устранения морфологической омонимии (многозначности) слов
 - Выбор правильной леммы
 - Уточнение морфологических характеристик

Основные методы

- Написание правил,
- Статистические методы, прежде всего, на основе морфологически размеченного корпуса

Примеры правил постморфологического анализа

- Удаление признаков служебных частей речи для однобуквенных слов, за которыми следуют точки
- Удаление омонимов слова «уже», соответствующих прилагательным, если за ним не стоит запятая или слово в родительном падеже
- Удаление омонимов слова «сорока», если после слова следует числительное (сорок пять)
- Обработка предлогов: удаление у слова, следующего за предлогом, всех омонимов, не соответствующих падежам, которыми обычно управляет данный предлог

Статистические методы и морфологическая разметка корпуса

Процедура морфологической разметки

- Морфологический анализ всех словоформ текста
- Снятие неоднозначностей (или исправление ошибок)
- Добавление информации о результатах в электронное представление текста
 - Лемма, части речи, грамматические характеристики

Фрагмент морфологической разметки в Национальном корпусе русского языка

- Я сидел на барском сиденье, дышал горячим ветром, бившим в лицо, ощущая в то же время не истребимую никакими сквозняками пыль и легкий запах духов -- катафалк с хорошей скоростью мчался по шоссе на юг.
(Ю. Трифонов)
- <s>**Я**{я=S,ед,од=им} **сидел**{сидеть=V,несов=изъяв,прош,ед,муж} **на**{на=PR} **барском**{барский=A=ед,сред,пр} **сиденье**{сиденье=S,сред,неод=ед,пр}, **дышал**{дышать=V,несов=изъяв,прош,ед,муж} **горячим**{горячий=A=ед,муж,твор} **ветром**{ветер=S,муж,неод=ед,твор}, **бившим**{бить=V,несов=прич,прош,ед,муж,твор} **в**{в=PR} **лицо**{лицо=S,сред,неод=ед,вин}, **ощущая**{ощущать=V=несов,деепр,непрош} **в**{в=PR} **то**{тот=A=ед,сред,вин} **же**{же=PART} **время**{время=S,сред,неод=ед,вин} **не**{не=PART} **истребимую**{истребимый=A=ед,жен,вин} **никакими**{никакой=A=мн,твор} **сквозняками**{сквозняк=S,муж,неод=мн,твор} **пыль**{пыль=S,жен,неод,ед=вин} **и**{и=CONJ} **легкий**{легкий=A=ед,муж,вин,неод} **запах**{запах=S,муж,неод=ед,вин}...

Проект OpenCorpora

[OpenCorpora](#)[Разметка ▾](#)[Словарь](#)[Статистика](#)[Скачать](#)[О проекте](#)[Бейджи](#)[Войти ▾](#)

Открытый корпус


Здравствуйте!

Это сайт проекта «Открытый корпус» (OpenCorpora). Наша цель – создать морфологически, синтаксически и семантически размеченный корпус текстов на русском языке, в полном объёме доступный для исследователей и редактируемый пользователями.

Мы начали работу в 2009 году, сейчас идёт разработка. Следить за тем, как мы продвигаемся, можно [здесь](#) (да, код проекта открыт).

Как я могу помочь прямо сейчас?

- принять участие в [разметке именованных сущностей](#) (см. [инструкцию](#))
- принять участие в снятии морфологической неоднозначности ([зарегистрируйтесь](#), чтобы получить доступ к заданиям, а также прочтите [руководство](#))
(всего мы получили уже **больше 2.96 млн** ответов)
- [предложить нам](#) источник свободно доступных (на условиях CC-BY-SA или совместимых) текстов
- добавить тексты в корпус (напишите нам письмо на opencorpora@opencorpora.org, мы расскажем как)
- помочь в разработке ПО корпуса и связанных с ним библиотек (тоже напишите нам письмо на opencorpora@opencorpora.org)
- рассказать о нас всем вокруг
- сделать ещё что-нибудь полезное и интересное (разумеется, напишите нам письмо на opencorpora@opencorpora.org)

 Мне нравитсяПонравилось **3102** людям НравитсяНравится 476 людям. [Зарегистрируйтесь](#), чтобы посмотреть, что нравится друзьям. Отзывы и предложения

[Разметка](#) / [именительный](#) — [винительный](#)

Спасибо, что помогаете нам. Не торопитесь, будьте внимательны. Если вы не уверены, пропускайте пример.

... , что пора в **главный** корпус , беспокоясь , ...

[именительный](#)[винительный](#)[Другое](#)[Пропустить](#)[Прокомментировать](#)

... на СС , но **этот** процесс долгий и сроки ...

[именительный](#)[винительный](#)[Другое](#)[Пропустить](#)[Прокомментировать](#)

... , на Украине предполагается **позтапное** повышение пенсионного возраста для ...

[именительный](#)[винительный](#)[Другое](#)[Пропустить](#)[Прокомментировать](#)

Если существует единое **эволюционное** дерево , объединяющее все ...

[именительный](#)[винительный](#)[Другое](#)[Пропустить](#)[Прокомментировать](#)

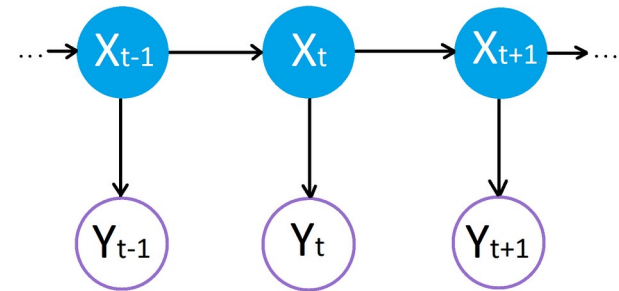
... , в котором за **мировое** первенство боролись сборные Испании ...

[именительный](#)[винительный](#)[Другое](#)[Пропустить](#)[Прокомментировать](#)[Хочу ещё примеров!](#)[Спасибо, достаточно](#)

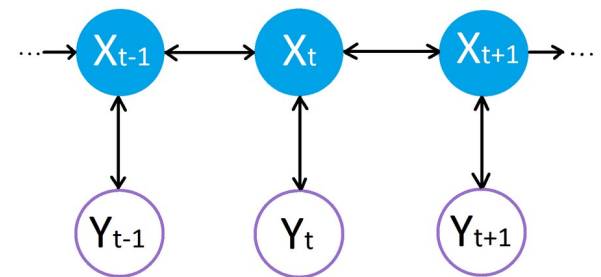
Методы автоматического разрешения морфологической неоднозначности

- Основная идея
 - Использовать контекст для предсказания правильного морф. тега

- Раньше
 - Скрытые марковские модели



- Метод CRF (conditional random fields) с 2003 года



- Сейчас нейронные сети

Морфологические процессоры русского языка

- АОТ (Диалинг) <http://www.aot.ru/>
 - На основе словаря, не поддерживается
- TreeTagger
 - Статистический анализатор на основе размеченного корпуса
 - Словарный
- PyMorphy2
 - Словарный, словарь проекта OpenCorpora
- MyStem (Yandex)
 - Словарный
- Snowball
 - стеммер

Сравнительные характеристики морф. анализаторов

Система	AOT	MyStem	TreeTagger	Rymorphy2
Открытые исходные коды	да	нет	нет	да
Скорость, слов в секунду	60-90 тыс.	100-120 тыс.	20-25 тыс.	80-100 тыс.
Подключение словарей	нет	да	да	нет
Объем словаря, тыс. слов	160	>250	210	250

Литература

- По морфологическому анализу и др. вопросам автоматической обработки
- https://miem.hse.ru/clschool/the_book
- Глава 2. Морфологический анализ текстов

Задание

- Установить морфологический анализатор:
 - Mystem
 - <http://company.yandex.ru/technology/mystem>
 - или pymorphy2
 - <https://pymorphy2.readthedocs.org/en/latest/>
- Создайте текстовую коллекцию из Ваших статей Википедии (все три запроса)
- Сделать частотные списки лемм из всех статей вместе
- Выдать по мере снижения частоты