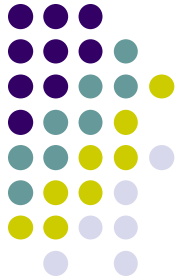
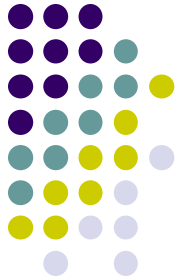


МОРФОЛОГИЧЕСКИЕ МОДЕЛИ И ПРОЦЕССОРЫ

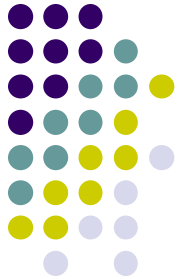
Большакова Елена Игоревна

СОДЕРЖАНИЕ



1. Основные понятия морфологии
   - Морфемика и Морфосинтаксис
   - Морфопараметры и морфопарадигма
   - Словоизменительные классы
   - Особенности русской морфологии
2. Морфологические процессоры
   - Функции процессоров
   - Словарные модели морфологии
   - Бессловарные модели морфологии
   - Машинное обучение для морфоанализа и разрешения морфологической омонимии
   - Морфопроцессоры для РЯ и др. ЕЯ
3. Заключение

ЭТАПЫ АНАЛИЗА ТЕКСТА В МНОГОУРОВНЕВЫХ МОДЕЛЯХ

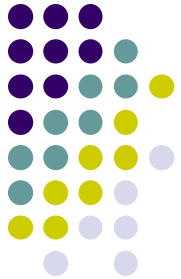


1. Предобработка: Графем. анализ / Сегментация
2. Морфологический анализ словоформы:
   – *лемматизация* (приведение к нормальной форме)
   – выявление морфопараметров
   – *стемминг* (получение основы слова)
3. Постморфологический анализ: разрешение (снятие) морфологической омонимии

➢ 1-3 – начальные этапы анализа

4. Синтаксический анализ – построение синтаксической структуры предложения
5. Семантический и дискурсивный анализ текста – построение семантического представления текста и определение его смысла

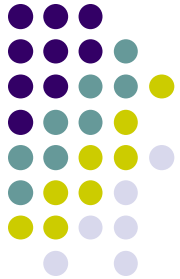
ПРЕДМЕТ И ФУНКЦИИ МОРФОЛОГИИ



- Морфология как раздел лингвистики изучает слова:
  - внутреннюю структуру (устройство) слов, (т.е. уровень морфем) – *морфемика*
  - внешние формы слов и их изменение в тексте (т.е. собственно уровень слов) – *морфосинтаксис*
- Словообразование (в языке) – создание новых слов
  - Словообразовательная парадигма:
    upgrade – retrograde – downgrade -…
    делать – сделать – недоделать – дело – деловой…
- Словоизменение (в тексте) – выражение нужной грамматической информации в тексте
  - Словоизменительная парадигма:
    стол – стола – столу – стол – столом – стола
    degrade – degrades – degrading – degraded

Парадигма (греч.) – образец, пример
В лингвистике: система форм слов

МОРФОЛОГИЯ: МОРФЕМИКА



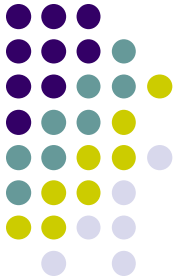
- Слова ЕЯ состоят из морфов (морфем)
  кус-ок пре-красн-ый happiness tree
- *Морфемы* – минимальные значащие единицы ЕЯ, возникли из слов: *hopeful*
  (в словоформах текста – *морфы)*
- Виды морфем
  - Корень (корневая морфема) – носитель основного компонента значения
  - Аффикс – служебная морфема (дополнит. смысл)
    префикс (приставка), *суффикс*,
    флексия (окончание), *постфикс* (частицы *ся, сь)*
    за-пых-а-вш-ий-ся
- Прикладные задачи: информационный поиск, распознавание неологизмов и родственных слов, исправление *паронимических* ошибок

МОРФОЛОГИЯ: МОРФОСИНТАКСИС

Словоизменение (изменение формы) –
согласно функции слов в составе предложений текста

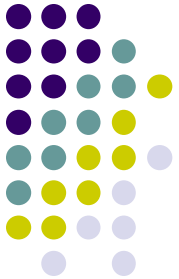
- Части речи, *POS (Part Of Speech)* – группы слов в зависимости от внешней формы и синтаксической функции в тексте (синтаксические классы), например: союзы
- Слова ЕЯ также разбиваются на:
  - знаменательные (самостоятельное лексическое значение): существительные, глаголы, прилагательные, наречия, причастия, числительные
  - служебные: предлоги, союзы, частицы (местоимения?)
- Группы слов с точки зрения словоизменения:
  - неизменяемые слова (служебные +наречия)
  - спрягаемые (глаголы) и склоняемые (имен. части речи)
- Грамматические формы слов в тексте зависят от *морфологических характеристик* (= параметров)

СЛОВО В ЯЗЫКЕ И ТЕКСТЕ



- *Словоформа* – конкретная грамматическая форма слова: *стола, забежал, eats*
- *Лексема* (единица словаря) – совокупность всех словоформ слова: {*стол, стола, столу, столом, столы, столов, столам, столами* }
  по сути, семантический инвариант
  Слово меняется в зависимости от его синт.функции
- *Лемма* – нормальная (базовая, каноническая), словарная форма (*имя лексемы*) : *стол, бросать*
  (для глаголов – инфинитив)
- Для анализа текста нужна лемма (лемматизация)
- *Основа слова* (*stem*) – часть слова без окончания и постфикса: *стол, стола* *кусок - куском - куска*
- *Псевдооснова* – неизменяемая начальная часть слова: *кусок - куска* *гнать - гонит*

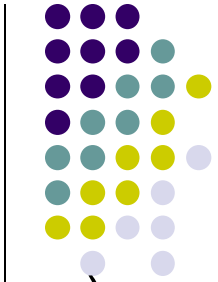
МОРФОЛОГИЧЕСКИЕ ПАРАМЕТРЫ



= *грамматические характеристики/категории* слов
= *морфологические признаки/параметры*
= *грамматические переменные* (имеют значения):

- Род (мужской, женский, средний, общий)
- Число (единственное, множественное)
- Одушевленность (да/нет)
- Падеж (от 1-3 до 16-18 в финском языке)
- Степень (сравнительная, превосходная)
- Лицо (первое, второе, третье)
- Время (настоящее, прошедшее, будущее)
- Вид (совершенный, несовершенный)
- Возвратность (да/нет)
- Залог (действительный, страдательный)
- Наклонение (изъявительное, сослагательное, повелительное)

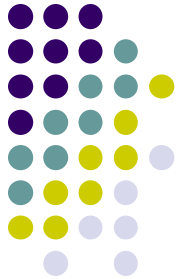
МОРФОПАРАМЕТРЫ: ЗНАЧЕНИЯ



- Значения морфологических параметров (МП):
  - связанные (фиксированные, словообразовательные), присущие лексеме в целом;
    например: род, одушевленность для существительных
  - свободные (изменяемые, формообразовательные), различны в разных словоформах лексемы;
    например: число, падеж существительных
- Часть речи (*Part of Speech, POS*) характеризует общий набор МП, в том числе свободных МП и их значений
- Набор свободных морфопараметров с конкретными значениями характеризует определенную словоформу

столов POS = Существительное,
Падеж = Родительный,
Число = Множественное,
(Род и Одушевленность фиксированы)

МОРФОЛОГИЧЕСКАЯ ПАРАДИГМА



- Морфологическая (= *словоизменительная*) парадигма – система форм слова:
  - изменение морфологических параметров слова
  - наличие инвариантной части (корень, основа)
  - перечень изменяемых компонентов (окончаний)
- Морфологическая парадигма слова *завод*:

  {*завод, завода, заводу, заводом, заводы, заводов, заводам, заводами*} ≡ парадигме слов *стол, вектор*

Единственное число						Множественное число					
Им.	Род	Дат	Вин	Тв.	Пр.	Им.	Род	Дат	Вин	Тв.	Пр.
ǿ	а	у	ǿ	ом	е	ы	ов	ам	ы	ами	ах

СЛОВОИЗМЕНИТЕЛЬНЫЙ КЛАСС

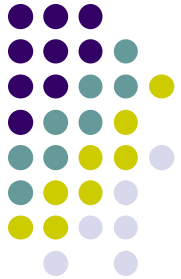
- В парадигме бывают исключения:
  – отсутствие некоторых форм: *ножницы, победить*
  – изменение основы: *идти – шел*
  (есть неизменяемые слова: *и*, *всего*, *пальто*)
- *Парадигматический/ Словоизменительный / Флективный класс*: подкласс слов одной части речи с одной морфопарадигмой: *{завод, стол, кот, …}* *стул* ?
- В морфологической модели с каждой лексемой/леммой связаны: часть речи (POS) и словоизменительный класс
- Число словоизменительных классов?
  различно для разных частей речи (изменяемых),
  зависит от модели морфологии
- Синтаксические подклассы (частей речи), например, в РЯ подклассы прилагательных: местоименные (*который*), притяжательные (*дядин*),порядковые числит. (*второй*)

ПРИМЕРЫ СЛОВОИЗМЕНИТЕЛЬНЫХ КЛАССОВ

Неодушевленные существительные мужского рода:

№	Слово (пример)	Падежные окончания	
		Един. число Им. Род. Дат. Вин. Твор. Пр.	Множ. число Им. Род. Дат. Вин. Твор. Пр.
01	Телефон	_ а у _ ом е	ы ов ам ы ами ах
02	Тираж	_ а у _ ом е	и ей ам и ами ах
03	Огонь	ь я ю ь ем е	и ей ям и ями ях
04	Перебой	й я ю й ем е	и ев ям и ями ях
05	Санаторий	й я ю й ем и	и ев ям и ями ях
…	…	…	…

ОСОБЕННОСТИ РУССКОЙ МОРФОЛОГИИ: ИМЕНА



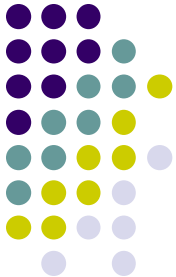
Существительные:

- падежи: 6 + 2 + 1 (есть совпадение форм в падежах):
  партитив (частичный): *рюмку коньяка /коньяку*
  локатив (местный): *в аэропорте /аэропорту*
- 2 числа, 3 рода + общий род: *ябеда, инвестор*
- форма вин. падежа одуш.сущ. м.рода ед.числа и одуш.сущ. мн.числа (любого рода) совпадает с ф. род. падежа : *оператор1*
- форма винит. падежа неодуш.сущ. м.рода ед.числа, всех сущ. сред.рода ед. числа, неодуш.сущ. мн.числа (любого рода) совпадает с формой именит. падежа: *оператор2*
- форма предложного падежа сущ. и прил. жен.рода ед.числа совпадает с формой дательного падежа: *белой*

Прилагательные:

- изменение по родам в единственном числе
- краткие формы: *плох, плоха, плохи*
- 4 степени сравнения:
  краснее, покраснее, краснейший, наикраснейший

ОСОБЕННОСТИ РУССКОЙ МОРФОЛОГИИ: ГЛАГОЛ



- Парадигма глагола: инфинитив + личные формы + причастия + деепричастия + формы совершенного и несовершенного вида
- Вид глагола: совершенный/несовершенный
  - Соверш. вид: нет форм наст. времени и страдательных форм
  - Несоверш. вид : аналит. форма будущ. времени: *буду писать*
- В наст. времени изменяются формы лица, в прошедшем – рода
- Непереходные глаголы не имеют возвратных форм и форм страдательного залога
- Причастия: парадигма совпадает с парадигмой прилагательных, но нет форм сравнительной степени
- Деепричастия: неизменяемые

В целом в морфологии РЯ много исключений, например:

– отсутствие некоторых форм в парадигме: *ножницы, рад*

– *беглые гласные* при изменении: *сон - сна, дубок – дубка*

СЛОВАРЬ ЗАЛИЗНЯКА

- Основа большинства словарных компьютерных моделей морфологии РЯ
- А.А. Зализняк, 1977 г. «Грамматический словарь русского языка»
  - 100 тыс. словарных входов (лексем)
  - системный подход к описанию морфологических парадигм РЯ, включающих не только изменение буквенного состава слов, но и ударения
  - парадигма глагола: личные формы + причастия + деепричастия
  - отражает (с помощью специальной системы условных обозначений) современное словоизменение РЯ
  - много старых слов (*при – пря*), нет новых (*вейпер*)

ФРАГМЕНТ СЛОВАРЯ ЗАЛИЗНЯКА

ТЕЧЬ

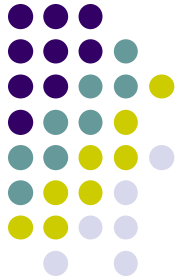
ж (жо): 8a, 8e, 8f″ — *47* | св (нсв): 8 — *118*

уте́чь св нп 8b/b (-к-), ё ◑II
вы́течь св нп 8a (-к-) ◑II
дичь ж 8a
на́взничь н
опри́чь предл.
стричь нсв 8b (-г-)
застри́чь св 8b (-г-) ◑II
настри́чь св 8b (-г-) ◑II
обстри́чь св 8b (-г-) ◑II
подстри́чь св 8b (-г-) ◑II
перестри́чь св 8b (-г-) ◑II
остри́чь св 8b (-г-) ◑II
достри́чь св 8b (-г-) ◑II
постри́чь св 8b (-г-) ◑II
простри́чь св 8b (-г-) ◑II
состри́чь св 8b (-г-) ◑II
расстри́чь св 8b (-г-) ◑II
отстри́чь св 8b (-г-) ◑II
вы́стричь св 8a (-г-) ◑II
засти́чь *см.* засти́гнуть
насти́чь *см.* насти́гнуть
пристичь *см.* пристигнуть
дости́чь *см.* дости́гнуть
пости́чь *см.* пости́гнуть
жёлчь ж 8a [// **желчь** =]

сво́лочь жо 8e
своло́чь св 8b/b (-к-) [// *простореч.* **сволочи́ть**] ◑I (-а-)
отволо́чь св 8b/b (-к-) [// *простореч.* **отволочи́ть**] ◑I (-а-)
уволо́чь св 8b/b (-к-) [// *простореч.* **уволочи́ть**] ◑I (-а-)
вы́волочь св 8a (-к-) [// *простореч.* **вы́волочить**] ◑I (-а-)
толо́чь нсв 8b/b (-к-) △ *наст.* толку́, толчёт, толку́т; *прош.* толо́к, толкла́, толо́кший; *прич. страд.* толчённый
затолочь св 8b/b (-к-) △ *буд.* затолку́, -чёт, -ку́т; *прош.* -о́к, -кла́, -о́кший; *прич. страд.* -чённый
натолочь св, *спряж. см.* затолочь
втолочь св, *спряж. см.* затолочь
подтолочь св, *спряж. см.* затолочь
перетолочь св, *спряж. см.* затолочь
потолочь св, *спряж. см.* затолочь
протолочь св, *спряж. см.* затолочь
столочь св, *спряж. см.* затолочь
растолочь св, *спряж. см.* затолочь

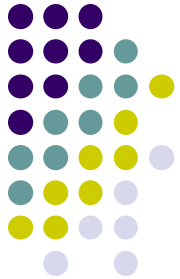
то́чь-в-то́чь н
запря́чь св 8b/b (-г-) ◑II
перезапря́чь св 8b/b (-г-) ◑II
напря́чь св 8b/b (-г-) ◑II
поднапря́чь св 8b/b (-г-)
перенапря́чь св 8b/b (-г-) ◑II
впрячь св 8b/b (-г-) ◑II
подпря́чь св 8b/b (-г-) ◑II
перепря́чь св 8b/b (-г-) ◑II
припря́чь св 8b/b (-г-) ◑II
сопря́чь св 8b/b (-г-) ◑II
спрячь св 8b/b (-г-) ◑II
распря́чь св 8b/b (-г-) ◑II
отпря́чь св 8b/b (-г-) ◑II
упря́чь св 8b/b (-г-) ◑II
вы́прячь св 8a (-г-) ◑II
на́отмашь н
ро́спашь ж 8a
гуа́шь ж 8a
плешь ж 8a
флешь ж 8a
брешь ж 8a
ишь част.; межд.
бишь част.
вишь част.
дишь (*часто без удар.*)

МОРФОЛОГИЧЕСКИЕ ПРОЦЕССОРЫ



- Морфопроцессоры могут выполнять:
  - Морфологический синтез нужной словоформы слова или всей его парадигмы исходя из леммы или основы
  - Морфологический анализ заданной словоформы:
    – приведение ее к лемме (*лемматизация*)
    – приведение к основе (*стемминг*)
    – анализ (выявление) морфопараметров словоформы
- Морфопроцессоры различаются способом создания:
  - На словарях и правилах, м.б. разная модель морфологии (способ представления лингвистической информации)
    - словарные морфологии
    - бессловарные морфологии
  - Машинное обучение на размеченном корпусе

МОРФОЛОГИЧЕСКИЙ СИНТЕЗ



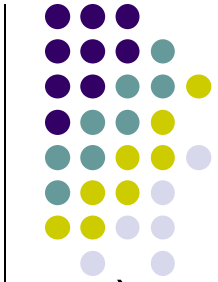
Два вида:

- Синтез нужной словоформы
  - Вход: лемма/основа + набор конкретных значений свободных морфологических параметров
  - Выход: словоформа:

  кипеть + изъяв.накл., прош.вр., ед.ч., м.род → *кипел*
- Синтез всей парадигмы слова (лексемы):
  - Вход: лемма/основа слова
  - Выход: словоизменительная парадигма:

  стол → стол, стола, столу, стол, столом, столе, столы, столов, столам, столы, столами, столов
- Возможен синтез разных вариантов словоформы (омонимия флексии): (*чашку*) *чая/чаю*

МОРФОЛОГИЧЕСКИЙ АНАЛИЗ: ЛЕММАТИЗАЦИЯ



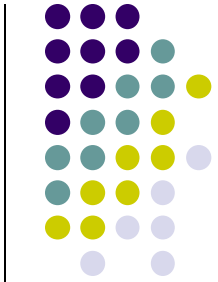
Переход от словоформы к её лемме (инварианту, лексеме)
т.е. приведение к словарной форме: *красивее → красивый*

- Лемматизация = *нормализация*
  - Вход: словоформа текста ЕЯ
  - Выход: лемма (иногда: + часть речи)

    лугового → луговой, луга → луг
- *Лемма* – словарная, обычно каноническая форма слова
  - для прилагательных РЯ –
    муж. род, ед. число, имен. падеж: *красивый*
  - для существительных – ед.число (?), имен. падеж: *стол*
  - для глагола? причастия? деепричастия

❖ Приложения: информационный поиск, классификация и кластеризация текстов в коллекциях

МОРФОЛОГИЧЕСКИЙ АНАЛИЗ: СТЕММИНГ

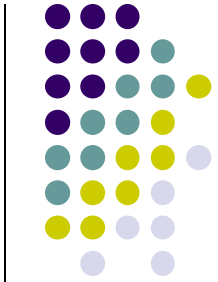


Приведение словоформы к основе/псевдооснове

- Вход: словоформа текста ЕЯ
- Выход: основа или псевдооснова:
  - *лесной, лесному, лесник → лесн*
  - *вода, водяного, водных → вод*

- Словоформы, соответствующие одной парадигме – словоизменительной (или словообразовательной) – должны получать одну и ту же (псевдо) основу
- Сложность: разные основы в парадигме (*идти* – *шел*)
- Стемминг в основном используется для текстов на малофлективных индоевропейских языках
- Приложения: информационный поиск, классификация и кластеризация текстов в коллекциях

СОБСТВЕННО МОРФОЛОГИЧЕСКИЙ АНАЛИЗ



Определение морфологических характеристик (*тегов*)

- Вход: словоформа текста ЕЯ
- Выход:
  - часть речи (*POS, part of speech*)
  - набор значений морфопараметров словоформы

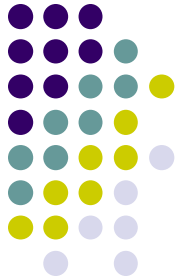
    водных → прил., мн.ч., род.п.

- Полный анализ – выявление всех морфохарактеристик
- *Тег* (*tag*) – запись морфологической характеристики

  Применяются разные системы тегов, зависят от ЕЯ

  UD (*Universal Dependency*) – универсальная система лингвистических тегов для многих ЕЯ
- Во многих приложениях нужен полный морфоанализ, но это зависит от конкретного языка

МОРФОЛОГИЧЕСКИЙ АНАЛИЗ: ТЕГИРОВАНИЕ



- Термин используется в основном в западной КЛ, буквально: проставление тегов к словоформам
- Чаще всего используется *part-of-speach (POS) tagging* , т.е. определение только части речи словоформы
  - Вход: словоформа текста ЕЯ
  - Выход: часть речи (*POS*)
    - *деревянного* → прил.
    - *composed* → *V* (verb) / *P* (participle)
- Применяется в основном для распространенных малофлективных индоевропейских языков, в частности, английского *Почему?*

❖ Приложения: разметка текста, например, перед синтаксическим анализом

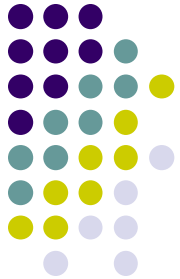
СЛОВАРНЫЕ МОРФОЛОГИИ

Компьютерные морфологические модели, основанные на одном из видов словарей:

- Словарь словоформ – каждой возможной словоформе приписана соответствующая морфологическая информация
  - объем словаря – число словоформ, зависит от степени флективности ЕЯ
- Словарь основ (или псевдооснов) – перечень существующих основ слов языка с морфологической информацией
  - объем словаря ~ число слов (лексем) в ЕЯ

- Основной морфологический словарь обычно дополняется более узкими словарями (окончаний, исключений и проч.)

МОДЕЛЬ НА ОСНОВЕ СЛОВАРЯ СЛОВОФОРМ

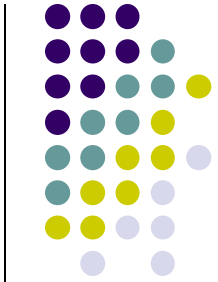


- Словарь содержит все словоформы учитываемых слов (лексем) с указанием для каждой их них:
  - морфол. характеристик словоформы, включая POS
  - леммы и/или основы
- Неизменяемые слова также обычно включаются в общий словарь
- Реализация морфоанализа: поиск заданной словоформы в словаре, в случае успеха выдается информация, приписанная найденной словоформе
- Быстрый поиск, но:
  - необходим существенный объем памяти:
    РЯ : словарь 100 тыс. слов – 2-2,5 млн. входов
  - трудоемкость построения и пополнения словаря

ФРАГМЕНТ СЛОВАРЯ СЛОВОФОРМ

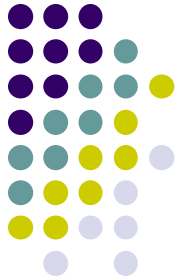
…	…	…	…	…
2609577	96056	одухотворяющие	ПРИЧ	дст, но, од, нст, им, мн
2609578	96056	одухотворяющих	ПРИЧ	дст, но, од, нст, рд, мн
2609579	96056	одухотворяющим	ПРИЧ	дст, но, од, нст, дт, мн
2609580	96056	одухотворяющих	ПРИЧ	дст, од, нст, вн, мн
2609581	96056	одухотворяющие	ПРИЧ	дст, но, нст, им, мн
2609582	96056	одухотворяющими	ПРИЧ	дст, но, од, нст, тв, мн
2609583	96056	одухотворяющих	ПРИЧ	дст, но, од, нст, пр, мн
2609584	96056	одухотворявший	ПРИЧ	дст, но, од, прш, мр, им, ед
2609585	96056	одухотворявшего	ПРИЧ	дст, но, од, прш, мр, рд, ед
2609586	96056	одухотворявшему	ПРИЧ	дст, но, од, прш, мр, дт, ед
2609587	96056	одухотворявшего	ПРИЧ	дст, од, прш, мр, вн, ед
…	…	…	…	…

МОДЕЛЬ НА ОСНОВЕ СЛОВАРЯ ОСНОВ



- Число словоизменительных классов на 3 порядка меньше числа словоформ и на 2 порядка – лексем,
  выгоднее хранить основы и отдельно – окончания ?
- Модель содержит словарь (псевдо)основ всех лексем, а также вспомогательные словари, в которых:
- Список окончаний (флексий/псевдофлексий) всех *словоизменительных* (флективных) *классов*
  - для каждой флексии указан набор значений морфологических характеристик, которые она может выражать (возможны варианты)
- Информация об особенностях словоизменения:
  - *чередования букв в основе*: *бегать – бежать*
  - *беглые гласные: кусок – куска, сон – сна, дуб – дубка*
  - *супплетивные формы*: *лучше – хороший, идти – шёл*

МОРФОАНАЛИЗ НА ОСНОВЕ СЛОВАРЯ ОСНОВ



Служебные и неизменяемые знаменательные слова могут храниться и обрабатываться отдельно.

Для остальных словоформ – обработка по схеме:

1. Отсекать последовательно возможные окончания длиной от 0 до 3 букв (в РЯ) , разбивая т.о. на основу и флексию;
2. Для полученного окончания проверить его наличие и определить его код по таблице окончаний, также найти номер его флективного класса по таблице классов;
3. Проверить наличие предполагаемой основы в словаре основ, в случае успеха найти номер ее флект. класса;
4. Проверить совпадение найденного флективного класса окончания и флективного класса предполагаемой основы;
5. В случае совпадения сохранить найденную информацию как результат морфоанализа;
6. Цикл отсечения окончаний продолжается и может быть выполнен даже в случае пустой псевдоосновы.

МОРФОАНАЛИЗ: ПРИМЕР

Анализ словоформы ***стола***

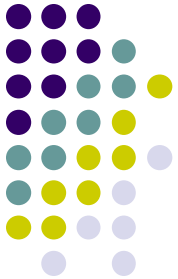
Основа	Номер флективного класса
…	…
СТОЛ	001
СТОЛБ	001
…	…

Номер флективного класса	Номер окончания	Номер морфолог. информации
…	…	…
001	20	36
001	22	40
001	66	06
002	67	26
002	66	17
…	…	…

Окончание	№
…	…
ят	62
ях	63
яя	64
_	65
а	66
е	67
и	70
…	…

№	Морфологическая информация
001	им. ед.
002	им. ед.; вн. ед.
003	им. ед.; вн. ед.; пр. ед.
004	им. ед.; вн. ед.; рд. мн.
005	им. ед.; рд. мн.; вн. мн.
006	рд. ед.
007	рд. ед.; дт. ед.; тв. ед.; пр. ед.
…	

ОСОБЕННОСТИ СЛОВАРНЫХ МОРФОЛОГИЙ



Представляют подход на правилах

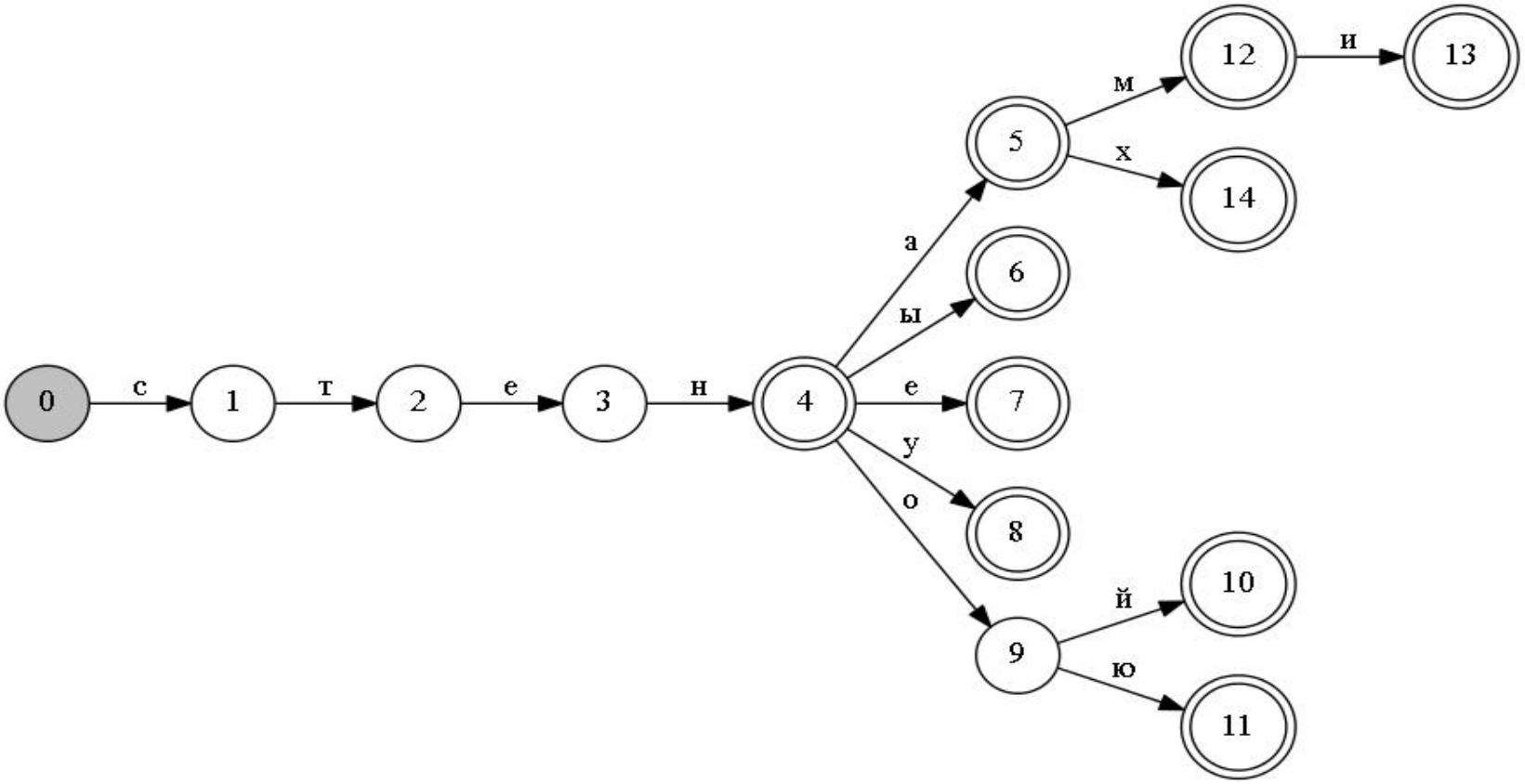
- Выполняют лемматизацию и полный морфоанализ
- Позволяют проводить синтез словоформ
- Дают возможность распознавания *морфологической омонимии:* *завод* – имен./вин.падеж?

Проблемы, требующие отдельного решения:

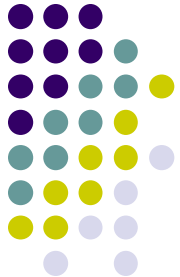
- Распознавание опечаток (Яндекс: ~480 опечаток в день в слове *одноклассники*; в месяц 1500 уник. опечаток)
- Анализ новых, *несловарных* слов (*out-of-vacabulary*), т.к. невозможен абсолютно полный словарь (лексика языка непрерывно пополняется)
- Требуется *разрешение морфологической омонимии*
- Для высокофлективных языков нужна оптимизация хранения словаря

МЕТОДЫ ХРАНЕНИЯ СЛОВАРЕЙ

- Хэширование (хэш-таблица и вспомогательная функция хэширования)
- Дерево словоформ, поиск выполняется конечным автоматом **DAWG** (*directed acyclic word graph*), скорость пропорциональна длине слова



АНАЛИЗ НЕСЛОВАРНЫХ СЛОВ В СЛОВАРНЫХ МОРФОЛОГИЯХ

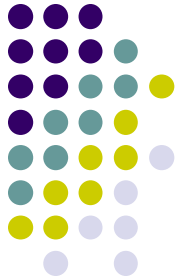


- Применяются эвристические методы предсказания морфологических параметров и леммы незнакомой словоформы – путем выявления аналогии со словоформами, представленными в словаре
- Предсказание по префиксу: отсечение префикса от анализируемой словоформы (не более 5 букв) и поиск словарной словоформы, совпадающей с остатком:
  ультраконсерватор – консерватор
- Предсказание по финали (окончанию) – поиск известной словоформы, которая имеет максимальное число общих конечных букв (2 и более) но анализируемое слово не вкладывается в нее: *зумить – курить, кринжовый* – ?
- В этих случаях словоформа разбирается по образцу так найденной словоформы, корректность анализа – 90-95%

БЕССЛОВАРНЫЕ МОРФОЛОГИИ

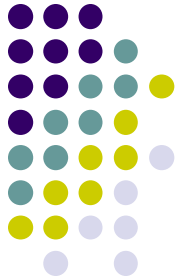
- Бессловарные морфологии: условно бессловарные, т.к. отсутствуют лишь большие словари (основ или словоформ), но фактически используется некоторая словарная информация:
  - словарь (таблица) псевдоокончаний (или аффиксов)
  - словарь неизменяемых слов (союзы, предлоги и др.)
  - правила преобразования словоформ, например, удаление/замена букв финали (окончания)
  - словарь исключений из правил
- Простейший вариант – определение части речи и других морфопараметров по финали (конечным буквам):
  -онок – сущ, муж.р., ед.ч, им.п. *мышонок позвонок*
- Для высокофлективных ЕЯ морфоанализ на основе бессловарных морфологий недостаточно точен
- Они успешно могут быть применены для стемминга

АЛГОРИТМЫ СТЕММИНГА



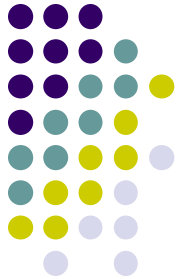
- Основаны на правилах преобразования словоформ, применяют списки служебных слов (предлоги, союзы, частицы, артикли и др.)
  - Пример правил для перевода существительных множ. числа в единственное (английский язык), замена окончаний:
    - *-ies → -y*
    - *-es → -ie*
    - *-s → ø*
- Исторически первый, для английского языка контекстно-зависимый алгоритм Ловинса (1968 г.) :
  - 294 окончания, каждое из которых может быть отсечено при выполнении некоторого условия
  - 35 правил трансформации слова после удаления окончания
  - выбор наиболее длинного окончания, которое можно отсечь при выполнении некоторых ограничений
  - ограничение: выделенная основа д.б. длиннее 3 символов
- Наиболее известный: **алгоритм Портера** (1980 г.) – развитие идей алгоритма Ловинса

АЛГОРИТМ ПОРТЕРА



- Достаточно быстрый алгоритм, 1980 г.
  изначально создан для английского языка,
  но разработаны модификации для европейских ЕЯ
- Приблизительно 60 правил преобразования входной словоформы, применяемых последовательно
- Правила имеют вид:
  <условие> <окончание> → <новое окончание>
- Например: *(m>0) E E D → E D*
  – если в словоформе есть хотя бы одна гласная и ее окончание *-eed*, то оно заменяется на *-ee* (*agreed → agree*)
- Ограничение:
  в основе слова должна остаться хотя бы одна гласная
- Используется специальная таблица исключений (например, для неправильных глаголов)
- Алгоритм реализован в системе *Snowball*, Для РЯ: http://snowball.tartarus.org/algorithms/russian/stemmer.html

ОСОБЕННОСТИ БЕССЛОВАРНЫХ МОРФОЛОГИЙ НА ПРАВИЛАХ



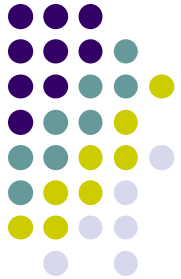
- Не зависят от объемных словарей
- Имеют хорошую скорость обработки словоформ
- Важно: дают возможность предсказания (с определенной вероятностью) практически любого нового слова

Проблемы:

- Чувствительны к коротким основам и их изменениям:
  шов, швы – основа *ш* ?
- Применимы в ограниченном круге приложений: даже если позволяют получать не только основу, но и лемму, она при этом фиксируется как новая, и потому нет передачи лексической информации на следующие уровни АОТ

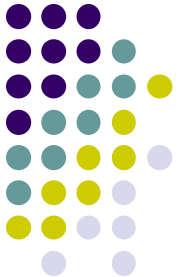
Но это касается и результатов обработки незнакомых слов в словарных моделях морфологиях.

МОРФОАНАЛИЗ: ВОПРОСЫ ДЛЯ ОСОЗНАНИЯ ТЕМЫ



- В чем отличие Словоформы от Токена?
- В чем отличие Леммы от Лексемы?
  Основы слова от Корня?
- Что входит в словоизменительную парадигму слова *прикладной ?*
- Каков должен быть результ лемматизации словоформ *спящей бежал бегал пропил ?*
- А результат полного анализа словоформы *зала* ?
- Что будет результатом стемминга словоформ *систему, системного, систематизировал, системами, …*
- Какой вид морфоанализа словоформ лучше для РЯ ?

РАЗРЕШЕНИЕ МОРФОЛОГИЧЕСКОЙ ОМОНИМИИ

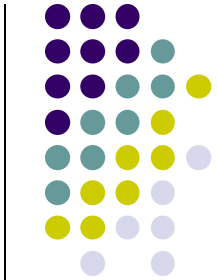


Разрешение / *Снятие* морфологической омонимии, *Morphological Disambiguation* – устранение многозначности:

- выбор правильной части речи и леммы
- уточнение морфологических характеристик

- Ранее реализовывалось в отдельном модуле (как постморфологический анализ), теперь эта функция встроена в большинство морфопроцессоров
- Снятие омонимии упрощает последующие этапы анализа:
  - сокращается объем хранимой информации
  - уменьшается количество деревьев разбора на этапе синтаксического анализа предложений
  - сокращается число различных значений слова при семантическом анализе
- Методы: Лингвист. правила или Машинное обучение

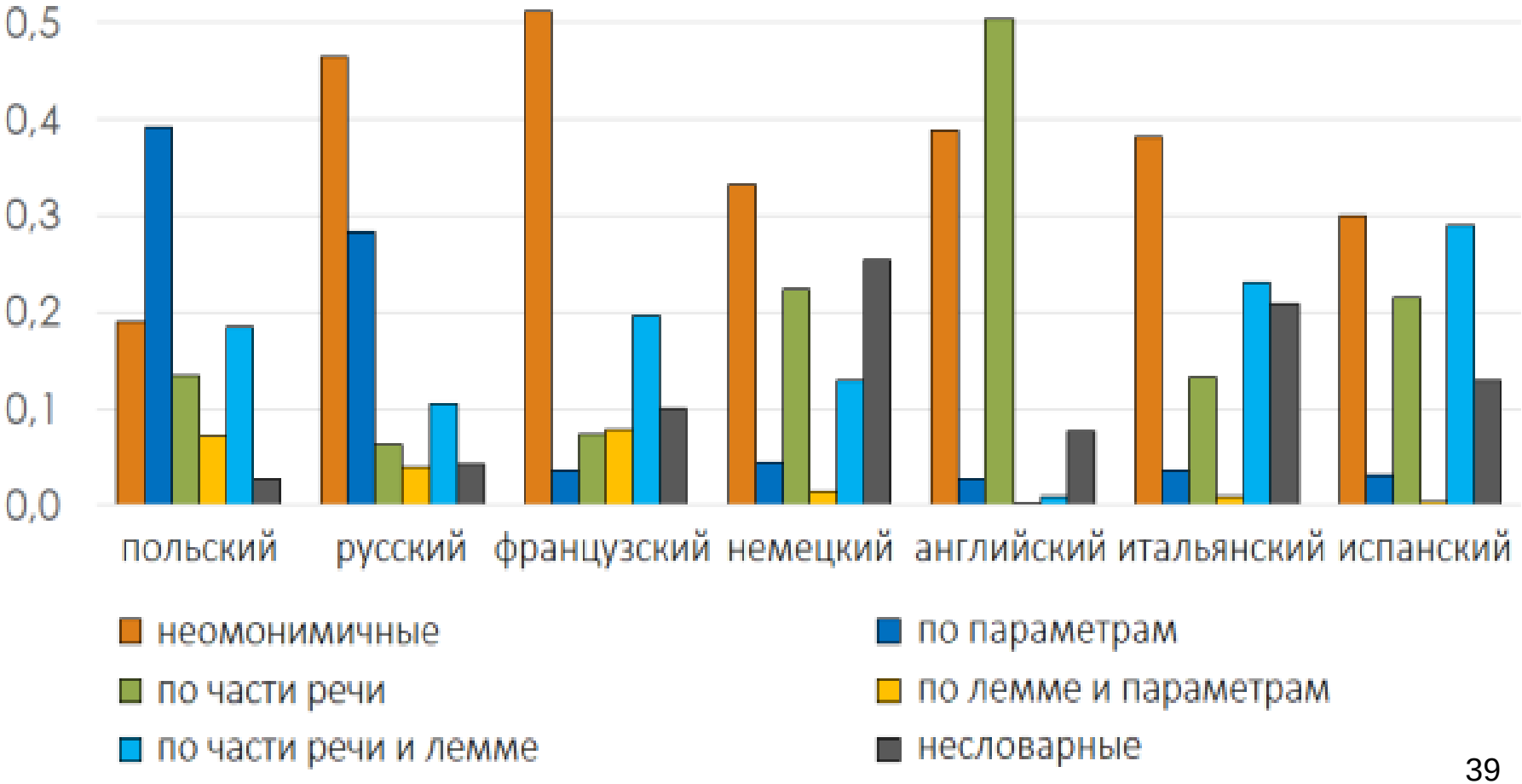
ВИДЫ МОРФОЛОГИЧЕСКОЙ ОМОНИМИИ



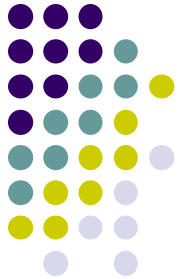
- *Лексико-морфологическая омонимия* – совпадение словоформ двух разных лексем:
  - *плачу* – 3 варианта:
    глагол *плакать* 1-го лица един. числа наст. времени
    глагол *платить* 1-го лица един. числа наст. времени
    существительное *плач* муж. рода, ед.ч., дат.падеж
  - *дома* – 2 вар-та: сущ., ед.ч., род.пад. + наречие (*д`ома*)
  - *стали* – 6 вариантов ? вина – ? вариантов
- Собственно *Морфологическая омонимия* – совпадение форм одного и того же слова (лексемы):
  телефон – 2 варианта: сущ., именит. и винит. падеж
- *Частеречная омонимия* – совпадение слов разных частей речи: *больной* – существит. и прилагательное
  а лемма при этом – одна ?

МОРФОЛОГИЧЕСКАЯ ОМОНИМИЯ: ОСТРОТА ПРОБЛЕМЫ

Распределение слов по классам омонимии
в текстах различных языков



РАЗРЕШЕНИЕ МОРФООМОНИМИИ НА ПРАВИЛАХ

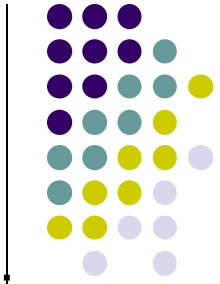


Лингвистические правила, примеры:

- Удаление омонимов слова с падежами, не соответствующими допустимым падежам предшествующего предлога:
  у зала – возможен предложный, но не именит. падеж
  (правило процессора Диалинг-АОТ)
- Выбор из двух вариантов: существительное или глагол – выбрать существительное, если до него стоит предлог: *посмотреть в стекло*
  – существительное *стекло*

Главный недостаток таких правил – очень низкая точность

МАШИННОЕ ОБУЧЕНИЕ ДЛЯ СНЯТИЯ МОРФОЛОГИЧЕСКОЙ ОМОНИМИИ



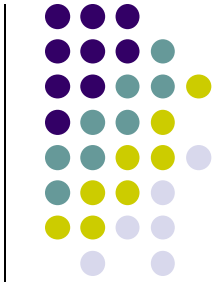
Несколько вариантов морфолог. разбора, необходимо:

- выбрать один (верный) или
- упорядочить варианты по степени вероятности

Возможные решения:

- Бесконтекстное снятие: подсчет по *размеченному корпусу* статистики разных вариантов разбора словоформы и их упорядочение (апостериорная вероятность вариантов)
- Контекстное разрешение: машинное обучение с учетом контекста (разборов словоформ до и после), обычно решается как *задача разметки последовательностей* (набор классификаторов для всех морфопраметров)
  - Скрытые марковские модели (HMM)
  - Случайные условные поля (CRF)
  - Нейронные сети RNN и BiRNN – до 97-98% точности

МОРФОЛОГИЧЕСКИЙ АНАЛИЗ НА БАЗЕ МАШИННОГО ОБУЧЕНИЯ

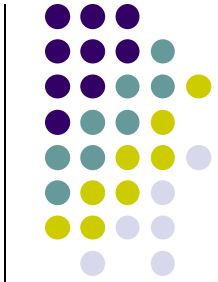


Цель: выполнять сразу анализ и снятие омонимии

Основа: большой размеченный текстовый корпус (в котором проставлены теги всех словоформ)

- 2016-2018 г. – Нейросетевые методы, с использованием рекуррентных сетей RNN и BiRNN, несколько слоев сети
- Вход сети: *эмбеддинги* (*embeddings*) – векторные представления слов (отображения слов в вектора веществ. чисел) от языковых моделей *Word2Vec, FastText*
- Часто: по отдельности определение морфологических характеристик, и для каждой характеристики – отдельный полносвязный слой сети (классификатор)
- Тогда последующее определение всего (наиболее вероятного) набора характеристик: например, логистической регрессией
- Лемматизация и определение POS – неплохо, хуже – полный морфоанализ для высоко флективных ЕЯ

МОРФОАНАЛИЗ НА НЕЙРОСЕТЯХ



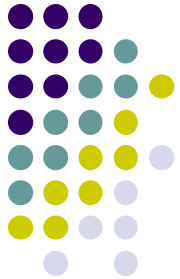
- С 2018-19 гг.– применение *контекстуализированных эмбеддингов* (векторных представлений слов,с учетом их контекста, от языковых моделей типа BERT)
- Определение морфохарактеристик и лемматизация словоформ проводятся по отдельности, независимо
- Лемматизация может рассматриваться как преобразование одной последовательности букв в другую (словоформы в лемму). Вход сети: – буквы обрабатываемой словоформы – ее морфологические характеристики

  Затем определение наиболее подходящего правила преобразования, автоматически по размеченному корпусу

Недостатки:

- Зависимость от обучающего корпуса: нет всех возможных комбинаций тегов (для корректного анализа редких слов)
- Производительность (по памяти и времени) на несколько порядков ниже, чем у методов на основе словарей

МОРФОПРОЦЕССОРЫ ДЛЯ РЯ



Свободный доступ, словарные морфологии

Но! имеют разные системы морфологических тегов и разные форматы вывода:

Морфоанализатор ***Mystem*** версии 3.0 компании Яндекс
http://company.yandex.ru/technology/mystem

- Токенизация всего текста
- Словарь: 200 тыс.лемм, ~20 МБ
- все функции морфологического анализа словоформ, в том числе разбор незнакомых слов
- два метода разрешения морфологической омонимии (контекстный, бесконтекстный), ранжирование вариантов
- нет морфологического синтеза
- закрытый код, вызов исполняемого модуля
- доступен в виде динамической библиотеки и позволяет подключать собственные словари

МОРФОПРОЦЕССОРЫ ДЛЯ РЯ: *ДИАЛИНГ-АОТ*

Первый открытый процессор для РЯ www.aot.ru,с 2004г.

- графематич.анализ, сегментация на предложения
- все функции морфологического анализа словоформ, на основе словаря А.А. Зализняка
- словарь: 174 тыс.лемм, ~ 9 МБ, можно пополнять
- обработка незнакомых слов по аналогии
- морфологический синтез полной парадигмы
- веб-интерфейс http://aot.ru/demo/morph.html
- открытый код на C++, подключение кода модуля

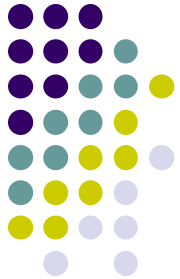
Делай

Found	Dict ID	Lemma	Grammems
+	пе, нс	ДЕЛАТЬ	Г дст,пвл,2л,ед

Печь

+	пе, нс	ПЕЧЬ	ИНФИНИТИВ, дст
+	но	ПЕЧЬ	С жр,вн,им,ед,

МОРФОПРОЦЕССОРЫ ДЛЯ РЯ: *Pymorphy2*

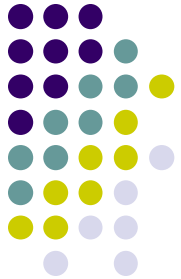


Переработка процессора Диалинг-АОТ для *Python*, автор – М. Коробов (с 2012-15 гг.)

- открытый код https://github.com/kmike/pymorphy2
- упрощенная токенизация
- все функции морфологического анализа словоформ
- словарь: 250 тыс. лемм, ~ 7МБ, формируется из словаря *OpenCorpora* с разметкой http://opencorpora.org/ возможно пополнение словаря
- только бесконтекстное снятие омонимии
- морфологический синтез
- кроме модулей для русского, есть для украинского
- документация pymorphy2.rtfd.org

https://pymorphy2.readthedocs.org/en/0.2/user/index.html

ДРУГИЕ МОРФОПРОЦЕССОРЫ



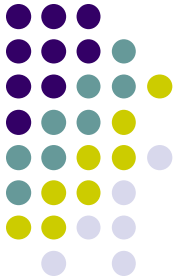
- Анализатор *TreeTagger* http://corpus.leeds.ac.uk/mocky/
  - закрытый код, вызов исполняемого модуля
  - только со снятой омонимией (машинное обучение)
- Морфопроцессор *CrossMorphy*
  - https://github.com/alesapin/XMorphy
  - открытый код на C++, морфологический синтез, теги *UD*
  - два метода разрешения морфологической омонимии
- Модуль стемминга *Snowball* (алгоритм Портера) для РЯ http://snowball.tartarus.org/algorithms/russian/stemmer.html
- Морфоанализатор для РЯ от *DeepPavlov,* нейросетевой, но лемматизация – от *Pymorphy2*
- Морфоанализаторы в рамках доступных многомодульных проектов (*pipelines*), например, *UDPipe*
  - *Natasha* https://github.com/natasha для русского языка
  - *Spacy* – библиотека с открытым кодом (на *Python*) для разных ЕЯ https://nlpub.ru/SpaCy

МОРФОПРОЦЕССОРЫ в *NLTK*

nltk – библиотека на *Python* для многих ЕЯ

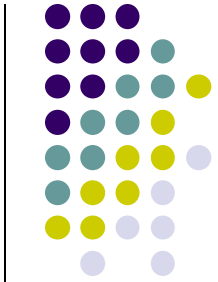
- 11 (?) морфологических процессоров в классе *stem*
- Модуль тегирования POS для английского языка
- Отдельные стеммеры: для английского (3), арабского (3), немецкого, португальского
- Стеммер для нескольких языков: Arabic, Danish, Dutch, English, Finnish, French, German, Hungarian, Italian, Norwegian, Portuguese, Romanian, Russian, Spanish, Swedish
- *WordNetLemmatizer* – лемматизация и тегирование для английского, основанные на словаре (единственный)

ПРИМЕРЫ РАБОТЫ *WordNetLemmatizer*



```
import nltk
from nltk.stem import WordNetLemmatizer

# Init the Wordnet Lemmatizer
lemmatizer = WordNetLemmatizer()

# Lemmatize Single Word
print(lemmatizer.lemmatize("bats")) #> bat
print(lemmatizer.lemmatize("feet")) #> foot
print(lemmatizer.lemmatize("are")) #> are

print(lemmatizer.lemmatize("are", 'v')) #> be

print(lemmatizer.lemmatize("stripes", 'v')) #>
strip
print(lemmatizer.lemmatize("stripes", 'n')) #>
stripe
```

ЗАКЛЮЧЕНИЕ



- Морфология охватывает широкий круг явлений, существенная проблема морфоанализА – омонимия
- Современные морфопроцессоры обычно включают дополнительные функции:
  - предсказание (интерпретация) незнакомых слов
  - разрешение морфологической омонимии
  - токенизация и сегментация на предложения
- Выбор процессора и компьютерной морфол. модели (словарная / бессловарная / машинное обучение), ее сложность может зависеть от решаемой прикладной задачи, вида текстов, конкретного ЕЯ
- Актуальны соревнования по оценке методов морфологич. анализа: Конференция http://www.dialog-21.ru/evaluation/ (РЯ)

СПАСИБО ЗА ВНИМАНИЕ!