

Математические модели в лингвистике

2. Компьютерная лингвистика: подходы, проблемы, задачи

Мати Пентус, Александр Пиперски,
Алексей Сорокин

МГУ им. М. В. Ломоносова,
межфакультетский курс,
осенний семестр 2015–2016 учебного года

Математические методы в лингвистике

- ▶ Решение практических задач:
компьютерная лингвистика (1-я часть семестра)
- ▶ Решение научных задач (2-я часть семестра)

Компьютерная лингвистика

- ▶ Разные понимания методов и предмета компьютерной лингвистики
- ▶ «Компьютерная лингвистика – 1» vs. «Компьютерная лингвистика – 2»
[Селегей 2012]

Компьютерная лингвистика – 1

- ▶ Формальная, полная и логически непротиворечивая теория языка, которая может использоваться для компьютерного анализа текстов

Компьютерная лингвистика – 2

- ▶ Технология и методология решения конкретных практических задач без претензий на общую теорию языка:
 - ▶ написание текста
 - ▶ машинный перевод
 - ▶ поиск и извлечение информации
 - ▶ естественно-языковые интерфейсы
 - ▶ ...

КЛ-1 и КЛ-2

- ▶ Расцвет КЛ-1: 1950-е – 1970-е годы
- ▶ Расцвет КЛ-2: 1990-е – 2000-е годы

Русское склонение: КЛ-1

секунда	ж	1а	слобода	ж	1f
мунда	п	0 (языки мунда)	вода	ж, 1d' // устар. 1f' ◇	
ерунда	ж	1b—		на воду // на воду;	
¹ рында	мо	<жо 1а> (истор.: оруженосец)		спуститься (уйти и т. п.) под воду [// под	
² рында	ж	1а (судовой колокол)		воду]; ходить (пойти)	
дурьнда	мо-жо	1а		по воду; вымыть	
гирлянда	ж	1а		в двух (трех и т. д.)	
ода	ж	1а		водах	
свобода	ж	1а	подвода	ж	1а
несвобода	ж	1а	воевода	мо	<жо 1а>
			пагода	ж	1а

Решаем задачу

- ▶ Задача с Московской традиционной олимпиады по лингвистике и математике для школьников
- ▶ Исходим из предположения, что морфологический анализатор построен

Алгоритм

- ▶ Привести все слова в заголовках статей и в запросах к начальной форме всеми возможными способами
- ▶ Исключить из рассмотрения стоп-слова (предлоги и союзы)
- ▶ Найти те заголовки статей, которые пересекаются с запросом хотя бы по одной начальной форме

Помятые брюки



помятые брюки — 2 млн ответов



Найти

 **С чем носить летние брюки? » Домашняя копилка**kak-sdelat.su > 4277-s-chem-nosit-letnie-**bryuki**.html ▾

Поиск

А **помните** брюки-бананы? ... Летние узенькие **брюки**-дудочки визуально сделают ваши ножки более длинными и стройными.

Картинки

Видео

 **Как выбрать брюки?**kak-vibrat.ru > article/76/kak-vybrat-**bryuki**.aspx ▾

Карты

И **помните**, чтобы **брюки** долго не теряли первоначальный вид, тщательно следите и ухаживайте за ними.

Маркет

Омонимия

- ▶ **Омонимия** — одно из наиболее неудобных свойств для автоматической обработки естественного языка
- ▶ В естественной жизни мы почти не замечаем омонимию

Омонимия в юморе

- ▶ *Женщина перед свадьбой говорит «Ты мой» и только после свадьбы уточняет, что именно мыть*
- ▶ *Из окна дуло. Штирлиц выстрелил, дуло исчезло*

[Санников 2002, 283–284]

Предложение из НКРЯ

- ▶ *В кодексе мой грех стоит три года
общего режима [Андрей Рубанов.
Сажайте, и вырастет (2005)]*
- ▶ Сколько слов здесь имеет
неоднозначный разбор?

Предложение из НКРЯ

- ▶ *мой, стоит, три*: непонятна начальная форма
- ▶ *грех, года, общего*: непонятны грамматические признаки
- ▶ *В, кодексе, режима*: однозначно

Вывод Mystem (1)

В{в=PR=|в=S,сокр=пр,мн|=S,сокр=пр,ед|=S,сокр=вин,мн|=S,сокр=вин,ед|=S,сокр=дат,мн|=S,сокр=дат,ед|=S,сокр=род,мн|=S,сокр=род,ед|=S,сокр=твор,мн|=S,сокр=твор,ед|=S,сокр=им,мн|=S,сокр=им,ед}
кодексе{кодекс=S,муж,неод=пр,ед}
мой{мой=APPO=вин,ед,муж,неод|=APPO=им,ед,муж|мыть=V,несов,пе=ед,пов,2-л}

Вывод Mystem (2)

грех{грех=S,муж,неод=вин,ед|=S,муж,неод
=им,ед|грех=ADV,прдк=}

стоит{стоять=V,несов,нп=непрош,ед,изъяв,
3-л|стоять=V,несов=непрош,ед,изъяв,3-л}

три{три=NUM=им|=NUM=вин,неод|тереть
=V,несов,пе=ед,пов,2-л}

года{год=S,муж,неод=вин,мн|=S,муж,неод
=род,ед|=S,муж,неод=им,мн}

Вывод Mystem (3)

общего{общий=A=вин,ед,полн,муж,од|=A=род,ед,полн,муж|=A=род,ед,полн,сред}
режима{режим=S,муж,неод=род,ед}

NB: даже чуть больше
неоднозначностей при разборе, чем мы
ожидали

Разрешение омонимии

Можно учесть:

- ▶ синтаксическую структуру (КЛ-1)
- ▶ контекст (КЛ-1 / КЛ-2)
- ▶ частотность (КЛ-2)

Разрешение омонимии

*В кодексе мой грех стоит три года
общего режима [Андрей Рубанов.
Сажайте, и вырастет (2005)]*

Как учесть:

- ▶ синтаксическую структуру?
- ▶ контекст?
- ▶ частотность?

Синонимия

- ▶ **Синонимия** — один и тот же объект может называться разными способами
- ▶ *бегемот = гиппопотам*
- ▶ Синонимичными бывают не только слова, но и словосочетания, предложения и тексты

Автоматическое извлечение информации

- ▶ Формализованное представление фактов в виде **Кто сделал что с кем где и когда?**
- ▶ *MergerBetween(company₁, company₂, date)*

Автоматическое извлечение информации

- ▶ *Представитель "Аэрофлота" подтвердил РИА Новости планы компании по покупке пакета акций "Трансаэро".*
- ▶ *ФАС получила ходатайство "Аэрофлота" о покупке "Трансаэро".*
- ▶ *«Аэрофлот» приобрел 75% плюс одну акцию авиакомпании «Трансаэро» за 1 рубль.*

Анафора и кореферентность

- ▶ **Анафора** — отсылка к ранее названной сущности с помощью местоимения
- ▶ **Кореферентность** — ситуация, когда две или более цепочки слов отсылают к одной и той же сущности
- ▶ Эти явления очень усложняют автоматическое извлечение информации из текстов

Анафора и кореферентность

Самая трагическая новость минувшей недели пришла из деревни Лука Новгородской области. Здесь сгорел психоневрологический интернат — из шестидесяти больных спаслись только двадцать три. Всех их вывела из горящего здания санитарка Юлия Ануфриева. Сама она погибла, спасая очередного пациента. Корреспондент «РР» отправилась в деревню Лука, чтобы узнать побольше об этом человеке.

Юля побежала в палату, услышав запах дыма. Угол палаты горел. Она бросилась тушить, но поняла, что это не помогает. <...>

Анафора и кореферентность

- ▶ *Юля побежала в палату, услышав запах дыма. Угол палаты горел. Она бросилась тушить*
- ▶ *Петя положил книгу на стол. Он был сделан из красного дерева.*

Спортивный репортаж

Предматчевые расклады для «Спартака» в восьмом туре были таковыми, что подопечные Дмитрия Аленичева могли подтянуться к призовой тройке. Ничья ЦСКА и «Зенита», случившаяся в субботу в Химках, и поражение «Локомотива» от «Рубина», давали шансы красно-белым на то, чтобы нагнать лидеров, сравнявшись по очкам с питерцами, и хозяева «Открытие Арены» намерены были воспользоваться такой возможностью. Однако, не всё у спартаковцев было так гладко

Спортивный репортаж

Предматчевые расклады для «Спартака» в восьмом туре были таковыми, что **подопечные Дмитрия Аленичева** могли подтянуться к призовой тройке. Ничья ЦСКА и «Зенита», случившаяся в субботу в Химках, и поражение «Локомотива» от «Рубина», давали шансы **красно-белым** на то, чтобы нагнать лидеров, сравнявшись по очкам с питерцами, и **хозяева «Открытие Арены»** намерены были воспользоваться такой возможностью. Однако, не всё у **спартаковцев** было так гладко

Прагматика. Знания о мире

- ▶ *Вы не знаете, который час?*
- ▶ *— Пойдешь со мной вечером в клуб?*
— У меня завтра утром экзамен.
- ▶ *Я видел их семью своими глазами.*
- ▶ Очень сложно моделируется как в КЛ-1, так и в КЛ-2

Задачи компьютерной лингвистики

- ▶ Распознавание текстов
- ▶ Распознавание и синтез речи
- ▶ Машинный перевод
- ▶ Проверка орфографии, грамматики и стиля
- ▶ Информационный поиск

Задачи компьютерной лингвистики

- ▶ Классификация и кластеризация текстов
- ▶ Фильтрация контента (спам и т. п.)
- ▶ Электронные словари
- ▶ Вопросно-ответные системы
- ▶ Естественно-языковые интерфейсы
- ▶ Извлечение информации и мнений



Санников В. З. 2002. *Русский язык в зеркале языковой игры*. Москва: Языки славянской культуры.



Селегей В. П. 2012. От автоматической обработки текста к машинному пониманию