

Математические модели в лингвистике

Вводная лекция, часть 2.

Мати Пентус, Александр Пиперски, Алексей Сорокин

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

Разделы лингвистики

- Разделы лингвистики:
 - Морфология,
 - Синтаксис,
 - Фонетика,

Разделы лингвистики

- Разделы лингвистики:
 - Морфология,
 - Синтаксис,
 - Фонетика,
 - Семантика,
 - Корпусная лингвистика,
 - ...,
- Как в них применяется математика?

Разделы лингвистики

- Разделы лингвистики:
 - Морфология,
 - Синтаксис,
 - Фонетика,
 - Семантика,
 - Корпусная лингвистика,
 - ...,
- Как в них применяется математика?
- И зачем нужны они сами в нелингвистических задачах?

Информационный поиск

- Одно из применений математической лингвистики — интернет-поиск.

Информационный поиск

- Одно из применений математической лингвистики — интернет-поиск.
- Казалось бы, лингвистика в поиске не нужна (просто ищем совпадение...),

Информационный поиск

- Одно из применений математической лингвистики — интернет-поиск.
- Казалось бы, лингвистика в поиске не нужна (просто ищем совпадение...),
- Однако всё не так просто:
 - Нужны не просто совпадения, а совпадения слов,

Информационный поиск

- Одно из применений математической лингвистики — интернет-поиск.
- Казалось бы, лингвистика в поиске не нужна (просто ищем совпадение...),
- Однако всё не так просто:
 - Нужны не просто совпадения, а совпадения слов,
 - Слова могут изменяться по грамматическим категориям,

Информационный поиск

- Одно из применений математической лингвистики — интернет-поиск.
- Казалось бы, лингвистика в поиске не нужна (просто ищем совпадение...),
- Однако всё не так просто:
 - Нужны не просто совпадения, а совпадения слов,
 - Слова могут изменяться по грамматическим категориям,
 - Пользователь может делать опечатки,

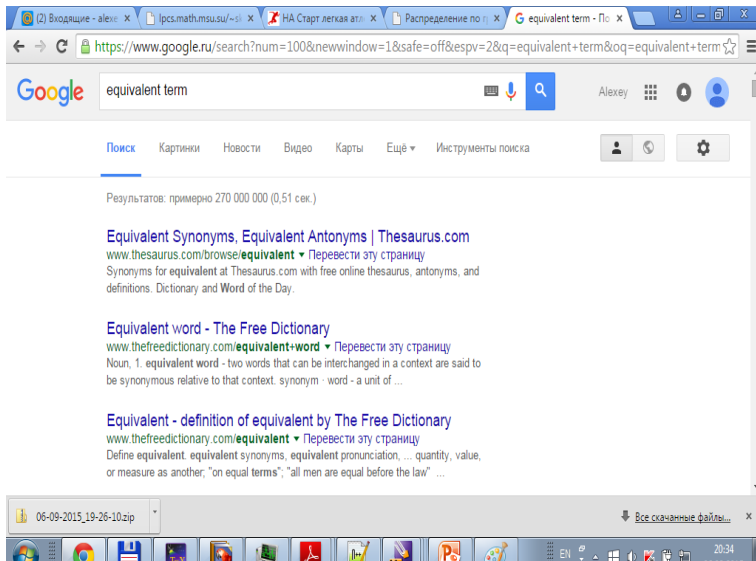
Информационный поиск

- Одно из применений математической лингвистики — интернет-поиск.
- Казалось бы, лингвистика в поиске не нужна (просто ищем совпадение...),
- Однако всё не так просто:
 - Нужны не просто совпадения, а совпадения слов,
 - Слова могут изменяться по грамматическим категориям,
 - Пользователь может делать опечатки,
 - Иногда пользователь не совсем знает, что ему нужно (стоит выдавать синонимы...),

Информационный поиск

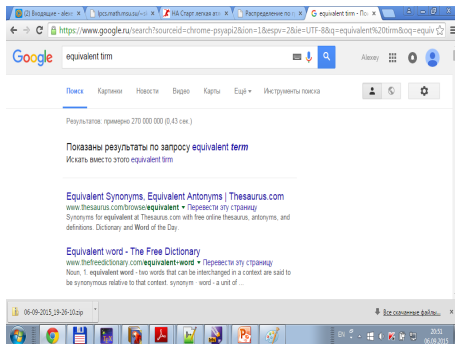
- Одно из применений математической лингвистики — интернет-поиск.
- Казалось бы, лингвистика в поиске не нужна (просто ищем совпадение...),
- Однако всё не так просто:
 - Нужны не просто совпадения, а совпадения слов,
 - Слова могут изменяться по грамматическим категориям,
 - Пользователь может делать опечатки,
 - Иногда пользователь не совсем знает, что ему нужно (стоит выдавать синонимы...),
 - Слова могут быть многозначны.

Пример поискового запроса



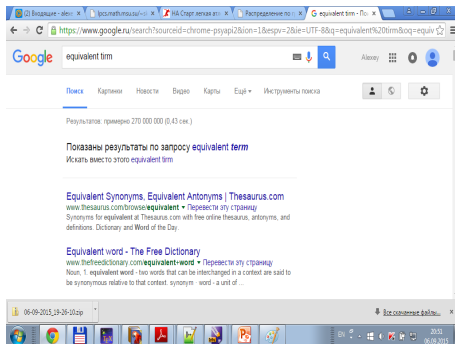
Выводы из примера

- Может быть найдено не только слово, но его форма (**term** $\mapsto$ **terms**),



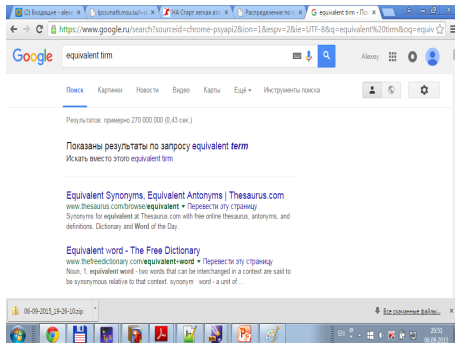
Выводы из примера

- Может быть найдено не только слово, но его форма (**term** $\mapsto$ **terms**),
- Может быть найдено синонимичное слово (**term** $\mapsto$ **word**),



Выводы из примера

- Может быть найдено не только слово, но его форма (**term** $\mapsto$ **terms**),
- Может быть найдено синонимичное слово (**term** $\mapsto$ **word**),
- Может быть исправлена опечатка (**tirm** $\mapsto$ **term**),



Вычислительная морфология

- Одна из задач вычислительной морфологии: восстановление форм слова по базовой форме,

Вычислительная морфология

- Одна из задач вычислительной морфологии: восстановление форм слова по базовой форме,

word+PI $\mapsto$ words,

state+PI $\mapsto$?,

Вычислительная морфология

- Одна из задач вычислительной морфологии: восстановление форм слова по базовой форме,

word+PI $\mapsto$ words,

state+PI $\mapsto$ states,

Вычислительная морфология

- Одна из задач вычислительной морфологии: восстановление форм слова по базовой форме,

word+PI $\mapsto$ words,

state+PI $\mapsto$ states,

- Однако не всё так просто:

Вычислительная морфология

- Одна из задач вычислительной морфологии: восстановление форм слова по базовой форме,

word + **PI** $\mapsto$ words,

state + **PI** $\mapsto$ **states**,

- Однако не всё так просто:

torch + **PI** $\mapsto$ torches,

Вычислительная морфология

- Одна из задач вычислительной морфологии: восстановление форм слова по базовой форме,

word + **PI** $\mapsto$ words,

state + **PI** $\mapsto$ **states**,

- Однако не всё так просто:

torch + **PI** $\mapsto$ torches,

daisy + **PI** $\mapsto$ daisies,

array + **PI** $\mapsto$ arrays,

Вычислительная морфология

- Одна из задач вычислительной морфологии: восстановление форм слова по базовой форме,

word+PI $\mapsto$ words,

state+PI $\mapsto$ states,

- Однако не всё так просто:

torch+PI $\mapsto$ torches,

daisy+PI $\mapsto$ daisies,

array+PI $\mapsto$ arrays,

mouse+PI $\mapsto$ mice,

goose+PI $\mapsto$ geese,

Вычислительная морфология

- Одна из задач вычислительной морфологии: восстановление форм слова по базовой форме,

word+PI $\mapsto$ words,
state+PI $\mapsto$ states,

- Однако не всё так просто:

torch+PI $\mapsto$ torches,
daisy+PI $\mapsto$ daisies,
array+PI $\mapsto$ arrays,
mouse+PI $\mapsto$ mice,
goose+PI $\mapsto$ geese,
cactus+PI $\mapsto$ cacti, cactuses,

Код для английского языка

```

### english.foma ###

read lexc irregular.lexc
define IrregularNounPlural;

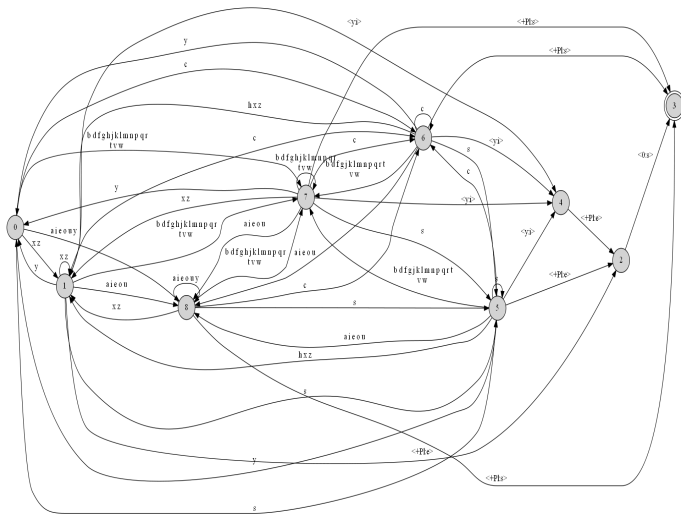
define Vowel [ a | i | e | o | u | y ];
define Consonant [ b | c | d | f | g | h | j | k | l | m | n | p | q | r | s | t |
    v | w | x | z ];
define Letter [ Vowel | Consonant ];
define Word [ Letter ]+;
define NounMark "+N";
define NounNumber "+Sg" | "+Pl";
define Noun Word NounMark NounNumber;

define NounAffixation "+N" "+Sg" -> "^" || _ .#.; "+N" "+Pl" -> "^" s || _ .#.;
define yReplacement y -> i e || Consonant _ "^" s .#.;
define sibException [ Letter ]+ a r c h _ "^" s ;
define Sibilant [ x | s | z | c h | s h ];
define elnsertion [..] -> e || Sibilant _ "^" s .#.;
define checkSibilant [ sibException .P. elnsertion ];
define Cleanup "^" -> [] || _ ;
define RegularNoun [ NounAffixation .o. yReplacement .o. checkSibilant .o. Cleanup
    ] ;
define Grammar Noun .o. [ IrregularNounPlural .P. RegularNoun ] ;

push Grammar

```


Конечный преобразователь для английского языка



Глагольные формы в языке йоулумни

основа	герундий	дуратив
saw “кричать”	saw-inay	sawaa-ʔaa-n
siim “разрушать”	siim-inay	siimii-ʔaa-n
hoyoo “называть”	hoy-inay	hoyoo-ʔaa-n
diiyl “охранять”	diyl-inay	diyiil-ʔaa-n
ʔilk “петь”	ʔilk-inay	ʔiliik-ʔaa-n
hiwiit “гулять”	hiwt-inay	hiwiit-ʔaa-n

Глагольные формы в языке йоулумни

Глагольные формы в языке йоулумни

основа	герундий	дуратив
saw “кричать”	saw-inay	sawaa-ʔaa-n
siim “разрушать”	siim-inay	siimii-ʔaa-n
hoyoo “называть”	hoy-inay	hoyoo-ʔaa-n
diiyl “охранять”	diyl-inay	diyiil-ʔaa-n
ʔilk “петь”	ʔilk-inay	ʔiliik-ʔaa-n
hiwiit “гулять”	hiwt-inay	hiwiit-ʔaa-n

Глагольные формы в языке йоулумни

Если основа имела вид $\alpha_1 V(V)\alpha_2(V)(V)\alpha_3$, где $\alpha_1, \alpha_2, \alpha_3 \in \{C, \varepsilon\}$, то основа герундия имеет вид $\alpha_1 V\alpha_2\alpha_3$,

Глагольные формы в языке йоулумни

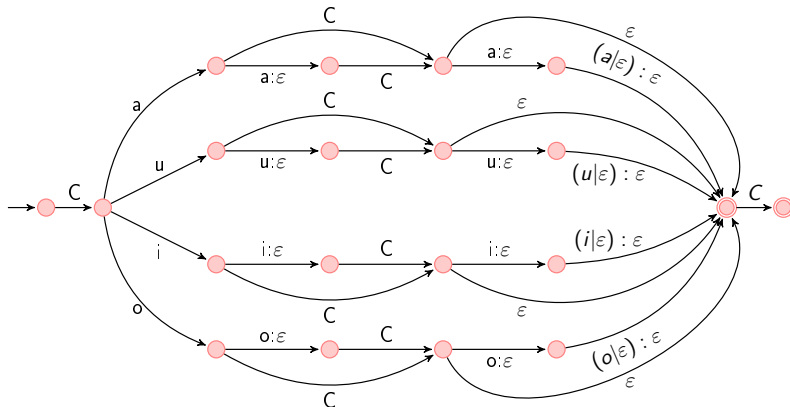
основа	герундий	дуратив
saw “кричать”	saw-inay	sawaa-ʔaa-n
siim “разрушать”	siim-inay	siimii-ʔaa-n
hoyoo “называть”	hoy-inay	hoyoo-ʔaa-n
diiyl “охранять”	diyl-inay	diyiil-ʔaa-n
ʔilk “петь”	ʔilk-inay	ʔiliik-ʔaa-n
hiwiit “гулять”	hiwt-inay	hiwiit-ʔaa-n

Глагольные формы в языке йоулумни

Если основа имела вид $\alpha_1 V(V)\alpha_2(V)(V)\alpha_3$, где $\alpha_1, \alpha_2, \alpha_3 \in \{C, \varepsilon\}$, то основа герундия имеет вид $\alpha_1 V\alpha_2\alpha_3$, а основа дуратива — $\alpha_1 V\alpha_2 VV\alpha_3$.

Преобразователи для глагольных форм в языке йоулумни

Основа герундия:



Постановка задачи

- Текстовая классификация — задача автоматического распределения текстов по классам.

Постановка задачи

- Текстовая классификация — задача автоматического распределения текстов по классам.
- Классы зависят от задачи:

Постановка задачи

- Текстовая классификация — задача автоматического распределения текстов по классам.
- Классы зависят от задачи:
 - Анализ тональности: положительные и отрицательные отзывы.

Постановка задачи

- Текстовая классификация — задача автоматического распределения текстов по классам.
- Классы зависят от задачи:
 - Анализ тональности: положительные и отрицательные отзывы.
 - Жанровая классификация: тексты различных жанров (новостные, энциклопедические, художественные...).

Постановка задачи

- Текстовая классификация — задача автоматического распределения текстов по классам.
- Классы зависят от задачи:
 - Анализ тональности: положительные и отрицательные отзывы.
 - Жанровая классификация: тексты различных жанров (новостные, энциклопедические, художественные...).
 - Тематическая классификация: распределение текстов по темам.

Постановка задачи

- Текстовая классификация — задача автоматического распределения текстов по классам.
- Классы зависят от задачи:
 - Анализ тональности: положительные и отрицательные отзывы.
 - Жанровая классификация: тексты различных жанров (новостные, энциклопедические, художественные...).
 - Тематическая классификация: распределение текстов по темам.
 - Языковая классификация: автоматический анализ языка.

Постановка задачи

- Текстовая классификация — задача автоматического распределения текстов по классам.
- Классы зависят от задачи:
 - Анализ тональности: положительные и отрицательные отзывы.
 - Жанровая классификация: тексты различных жанров (новостные, энциклопедические, художественные...).
 - Тематическая классификация: распределение текстов по темам.
 - Языковая классификация: автоматический анализ языка.
 - Можно классифицировать по полу, возрасту, социальному положению автора...

Постановка задачи

- Текстовая классификация — задача автоматического распределения текстов по классам.
- Классы зависят от задачи:
 - Анализ тональности: положительные и отрицательные отзывы.
 - Жанровая классификация: тексты различных жанров (новостные, энциклопедические, художественные...).
 - Тематическая классификация: распределение текстов по темам.
 - Языковая классификация: автоматический анализ языка.
 - Можно классифицировать по полу, возрасту, социальному положению автора...
- Применения:
 - Информационный поиск,

Постановка задачи

- Текстовая классификация — задача автоматического распределения текстов по классам.
- Классы зависят от задачи:
 - Анализ тональности: положительные и отрицательные отзывы.
 - Жанровая классификация: тексты различных жанров (новостные, энциклопедические, художественные...).
 - Тематическая классификация: распределение текстов по темам.
 - Языковая классификация: автоматический анализ языка.
 - Можно классифицировать по полу, возрасту, социальному положению автора...
- Применения:
 - Информационный поиск,
 - Автоматическая рубрикация,

Постановка задачи

- Текстовая классификация — задача автоматического распределения текстов по классам.
- Классы зависят от задачи:
 - Анализ тональности: положительные и отрицательные отзывы.
 - Жанровая классификация: тексты различных жанров (новостные, энциклопедические, художественные...).
 - Тематическая классификация: распределение текстов по темам.
 - Языковая классификация: автоматический анализ языка.
 - Можно классифицировать по полу, возрасту, социальному положению автора...
- Применения:
 - Информационный поиск,
 - Автоматическая рубрикация,
 - Рекомендационные системы.

Постановка задачи

- Текстовая классификация — задача автоматического распределения текстов по классам.
- Классы зависят от задачи:
 - Анализ тональности: положительные и отрицательные отзывы.
 - Жанровая классификация: тексты различных жанров (новостные, энциклопедические, художественные...).
 - Тематическая классификация: распределение текстов по темам.
 - Языковая классификация: автоматический анализ языка.
 - Можно классифицировать по полу, возрасту, социальному положению автора...
- Применения:
 - Информационный поиск,
 - Автоматическая рубрикация,
 - Рекомендационные системы.
 - Автоматическое определение авторства.

Модель “мешка слов”

- Для каждого слова учитывается только число его вхождений в текст.

Модель “мешка слов”

- Для каждого слова учитывается только число его вхождений в текст.
- Порядок слов не играет роли.

Модель “мешка слов”

- Для каждого слова учитывается только число его вхождений в текст.
- Порядок слов не играет роли.
- Получаем матрицу (таблицу) “объекты-признаки”.

Модель “мешка слов”

- Для каждого слова учитывается только число его вхождений в текст.
- Порядок слов не играет роли.
- Получаем матрицу (таблицу) “объекты-признаки”.
- Документ — строка матрицы (вектор значений признаков).

Модель “мешка слов”

- Для каждого слова учитывается только число его вхождений в текст.
- Порядок слов не играет роли.
- Получаем матрицу (таблицу) “объекты-признаки”.
- Документ — строка матрицы (вектор значений признаков).

документ	office	web	...	learn	метка класса
1	2	2	...	0	student
5	0	2	...	3	project
48	2	4	...	1	course

Пример матрицы “объекты-признаки” (WebKB corpus, McCallum, 1998)

Модель “мешка слов”

- Для каждого слова учитывается только число его вхождений в текст.
- Порядок слов не играет роли.
- Получаем матрицу (таблицу) “объекты-признаки”.
- Документ — строка матрицы (вектор значений признаков).

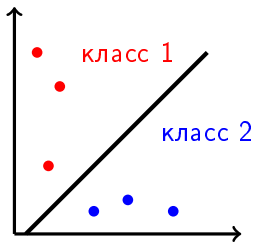
документ	office	web	...	learn	метка класса
1	2	2	...	0	student
5	0	2	...	3	project
48	2	4	...	1	course

Пример матрицы “объекты-признаки” (WebKB corpus, McCallum, 1998)

- Метод простой, но он работает во многих задачах.

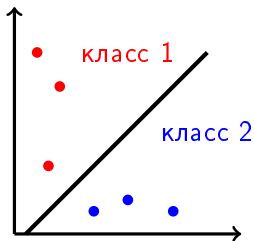
Особенности текстовой классификации

- Разделяющая поверхность — гиперплоскость



Особенности текстовой классификации

- Разделяющая поверхность — гиперплоскость

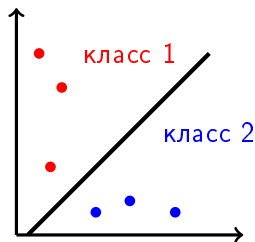


- Классификатор можно задать формулой:

$$h(x) = \text{sgn} \left(\sum_{i=1}^n w_i x_i - w_0 \right), \quad x_i \text{ — компоненты вектора } x$$

Особенности текстовой классификации

- Разделяющая поверхность — гиперплоскость



- Классификатор можно задать формулой:

$$h(x) = \text{sgn} \left(\sum_{i=1}^n w_i x_i - w_0 \right), \quad x_i \text{ — компоненты вектора } x$$

- Веса $w_0, w_1, \dots, w_n$ подбираются автоматически по обучающей выборке (текстам, для которых известна метка класса).

Особенности текстовой классификации

- Признаки могут учитывать контекст (биграммы, триграммы и т.д.),

Особенности текстовой классификации

- Признаки могут учитывать контекст (биграммы, триграммы и т.д.),
- В качестве признаков можно использовать длину предложения, статистику по грамматическим категориям, синтаксические показатели...

Особенности текстовой классификации

- Признаки могут учитывать контекст (биграммы, триграммы и т.д.),
- В качестве признаков можно использовать длину предложения, статистику по грамматическим категориям, синтаксические показатели...
- Подбор признаков — задача лингвиста.

Особенности текстовой классификации

- Признаки могут учитывать контекст (биграммы, триграммы и т.д.),
- В качестве признаков можно использовать длину предложения, статистику по грамматическим категориям, синтаксические показатели...
- Подбор признаков — задача лингвиста.
- Оптимальная классификация признаков — задача математика.

Особенности текстовой классификации

- Признаки могут учитывать контекст (биграммы, триграммы и т.д.),
- В качестве признаков можно использовать длину предложения, статистику по грамматическим категориям, синтаксические показатели...
- Подбор признаков — задача лингвиста.
- Оптимальная классификация признаков — задача математика.
- Трудности: большой объём данных, несбалансированная выборка, большая размерность данных, большое количество классов.