Математические модели в лингвистике

Вводная лекция, часть 2.

Мати Пентус, Александр Пиперски, Алексей Сорокин

МГУ им. М. В. Ломоносова, межфакультетский курс,
осенний семестр 2017–2018 учебного года,
12 сентября 2017 г.

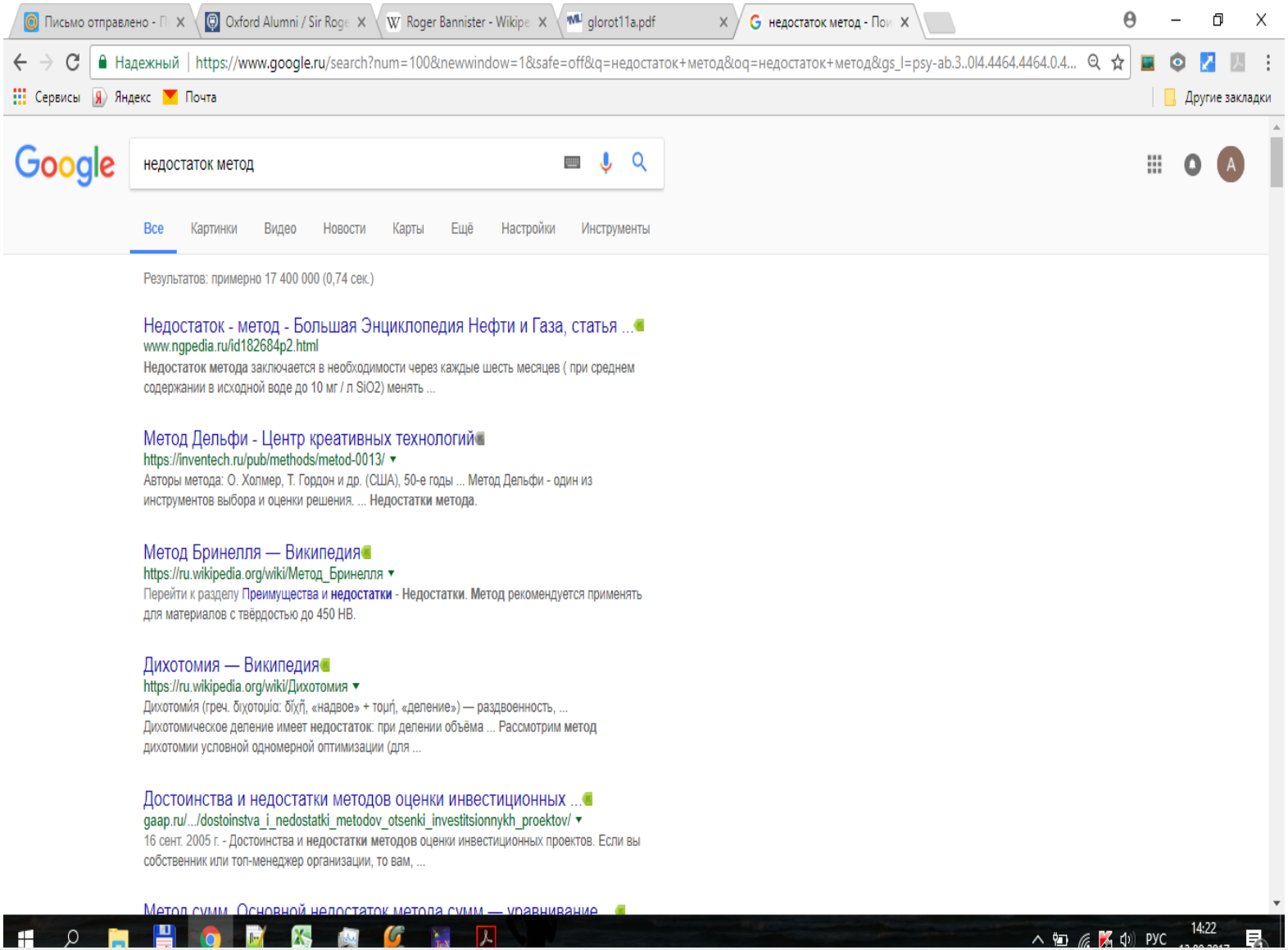
Разделы лингвистики

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

Информационный поиск

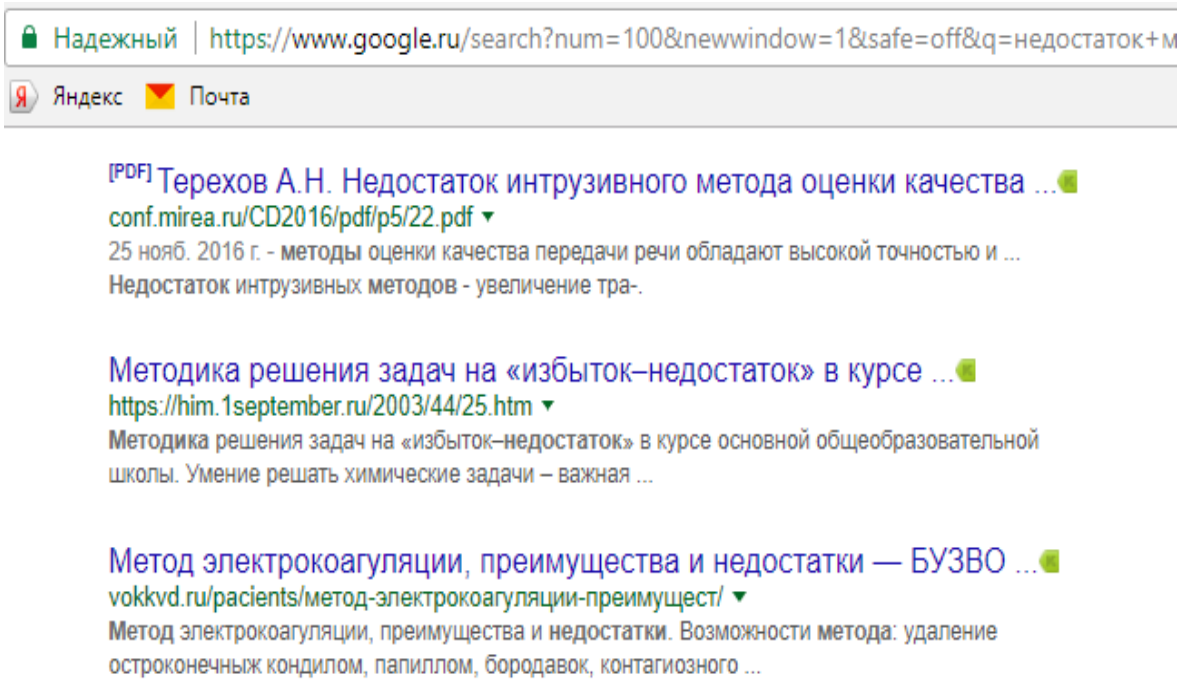
- Одно из применений математической лингвистики — интернет-поиск.
- Казалось бы, лингвистика в поиске не нужна (просто ищем совпадение...),
- Однако всё не так просто:
  - Нужны не просто совпадения, а совпадения слов,
  - Слова могут изменяться по грамматическим категориям,
  - Пользователь может делать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

Пример поискового запроса



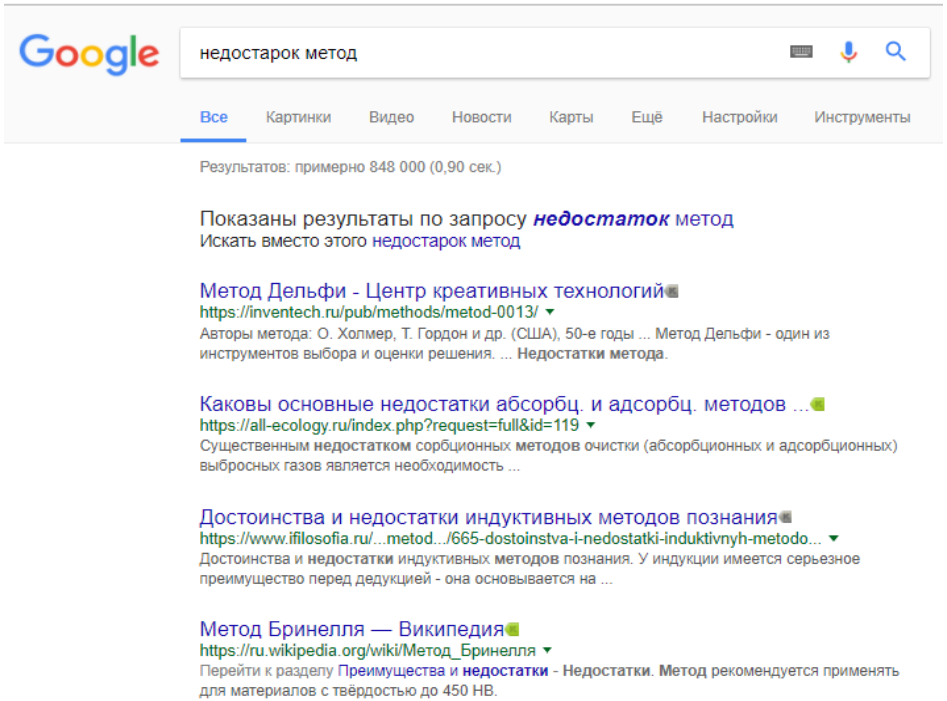
Выводы из примера

- Может быть найдено не только слово, но его форма(**метод** $\mapsto$ **метода**),
- Может быть найдено близкое по смыслу слово (**метод** $\mapsto$ **методика**),



Выводы из примера

- Может быть найдено не только слово, но его форма(**метод** $\mapsto$ **метода**),
- Может быть найдено близкое по смыслу слово (**метод** $\mapsto$ **методика**),
- Может быть исправлена опечатка (**недостарок** $\mapsto$ **недостаток**),



Вычислительная морфология

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

$$\begin{aligned} \text{word}+\textbf{Pl} &\mapsto \text{words}, \\ \text{state}+\textbf{Pl} &\mapsto ?, \end{aligned}$$

$$\begin{aligned} \text{word}+\textbf{Pl} &\mapsto \text{words}, \\ \text{state}+\textbf{Pl} &\mapsto \text{states}, \end{aligned}$$

- Однако не всё так просто:

$$\begin{aligned} \text{torch}+\textbf{Pl} &\mapsto \text{torch}\textbf{es}, \\ \text{daisy}+\textbf{Pl} &\mapsto \text{daisi}\textbf{es}, \\ \text{array}+\textbf{Pl} &\mapsto \text{array}\textbf{s}, \\ \text{mouse}+\textbf{Pl} &\mapsto \text{mice}, \\ \text{goose}+\textbf{Pl} &\mapsto \text{geese}, \\ \text{cactus}+\textbf{Pl} &\mapsto \text{cacti, cactuses}, \end{aligned}$$

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 eInsertion [..] -> e || Sibilant _ "^" s .#.;
define checkSibilant [ sibException .P. eInsertion ];
define Cleanup "^" -> [] || _ ;
define RegularNoun [ NounAffixation .o. yReplacement .o. checkSibilant .o. Cleanup
    ] ;
define Grammar Noun .o. [ IrregularNounPlural .P. RegularNoun ];

push Grammar
```

Русская глагольная морфология

Для русского всё куда хуже:

играть	$\mapsto$	*играет*
писать	$\mapsto$	*пишет*
спать	$\mapsto$	*спит*
греть	$\mapsto$	*греет*
хрипеть	$\mapsto$	*хрипит*
петь	$\mapsto$	*поёт*
целовать	$\mapsto$	*целует*

Глагольные формы в языке йоулумни

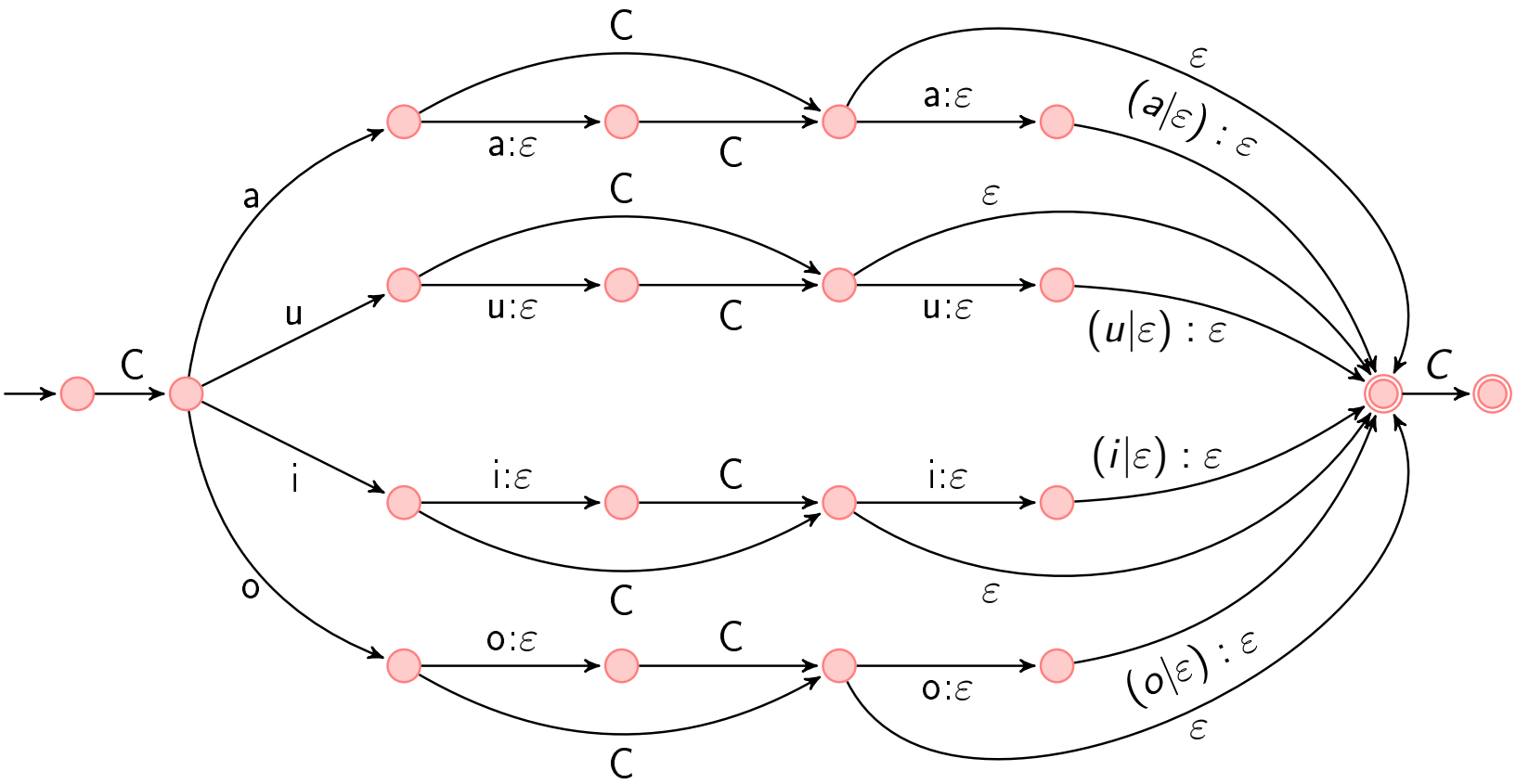
основа	герундий	дуратив
caw “кричать”	caw-inay	cawaa-ʔaa-n
cuum “разрушать”	cum-inay	cumuu-ʔaa-n
hoyoo “называть”	hoy-inay	hoyoo-ʔaa-n
diiyl “охранять”	diyl-inay	diyiil-ʔaa-n
ʔilk “петь”	ʔilk-inay	ʔiliik-ʔaa-n
hiwiit “гулять”	hiwt-inay	hiwiit-ʔaa-n

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.
- Классы зависят от задачи:
  - Анализ тональности: положительные, отрицательные и нейтральные 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.
- Порядок слов не играет роли.
- Получаем матрицу (таблицу) “объекты-признаки”.
- Документ — строка матрицы (вектор значений признаков).

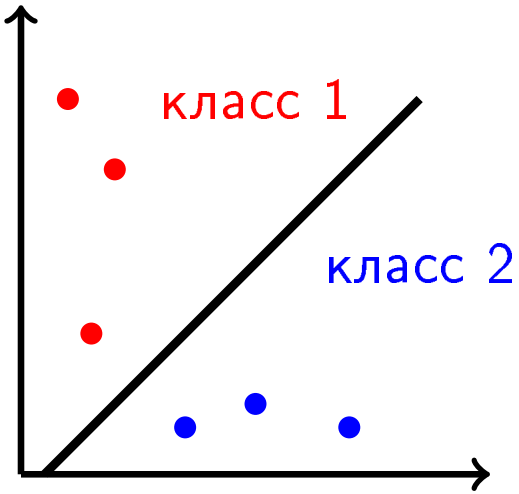
документ	office	web	...	learn	метка класса
1	2	2	...	0	student
5	0	2	...	3	project
48	2	4	...	1	course

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

- Метод простой, но он работает во многих задачах.

Особенности текстовой классификации

- Разделяющая поверхность — гиперплоскость



- Классификатор можно задать формулой:

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

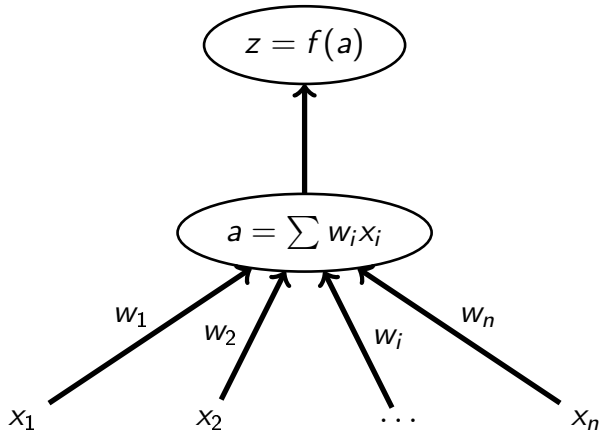
- Веса $w_0, w_1, \ldots, w_n$ подбираются автоматически по обучающей выборке (текстам, для которых известна метка класса).

Особенности текстовой классификации

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

Нейронные сети

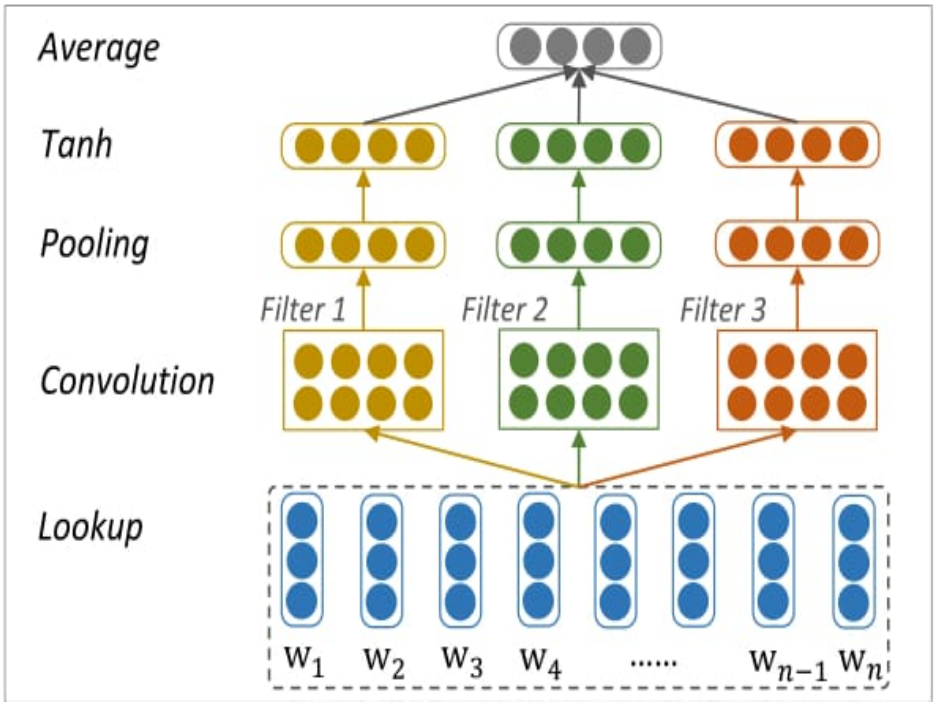
- Один из основных подходов: нейронные сети.
- Структура нейрона:



- Нейрон получает несколько сигналов $x_1, \dots, x_n$ и вычисляет их взвешенную сумму.
- Эта сумма поступает в функцию активации f.
- Варианты для функции f:
  - $f(x) = x$ (тождественная).
  - $f(x) = \operatorname{sgn}(x)$ (преодолён ли порог активации).
  - $f(x) = \dfrac{1}{1 + e^{-kx}}$ (преобразует $(-\infty; \infty)$ в $[0; 1]$).
  - $f(x) = max(x - b, 0)$ (ReLU, смесь тождественной и пороговой).

Нейронные сети

Реальные нейронные сети: десятки тысяч нейронов, органи-зованных в слои.



Структура сети и признаки зависят от лингвистической задачи (но не так сильно, как при классификации).