



# **Семантическая близость и разрешение неоднозначностей**

Евгений Борисов

# Семантическая близость и разрешение неоднозначностей

## Уровни сложности при автоматической обработке текстов

**Прагматика (Дискурс)** - смысловые контексты

**Семантика** - смыслы последовательностей слов

**Синтаксис** - правила формирования последовательностей слов

**Лексика** - отдельные слова и устойчивые словосочетания

# Семантическая близость и разрешение неоднозначностей

## Семантика

- лексическая, отдельные слова
- композиционная, комбинации слов

### задачи:

- разрешение многозначности
- оценка семантической близости

# Семантическая близость и разрешение неоднозначностей

## Неоднозначности в языке

омонимия - случайное совпадение слов

ключ, лук, замок, печь

полисемия - несколько связанных значений

стол <организация или объект> , платформа <политическая или железнодорожная>

метонимия - замена смысла

Целых три тарелки съел.

# Семантическая близость и разрешение неоднозначностей

## Отношения между словами

синонимия - общий смысл

машина, автомобиль

антонимия - противоположность

большой / маленький, вверх / вниз

гипонимия - обобщение

яблоко / фрукт, овчарка / собака

партономия - часть, вхождение

колесо / автомобиль, житель / город

# Семантическая близость и разрешение неоднозначностей

## WordNet

- База лексических отношений
  - содержит иерархии
  - сочетает в себе тезаурус и словарь
  - доступен on-line
  - разрабатываются версии для языков кроме английского (в т.ч. для русского)

Категория	Уникальных форм
Существительные	117,097
Глаголы	11,488
Прилагательные	22,141
Наречия	4,601

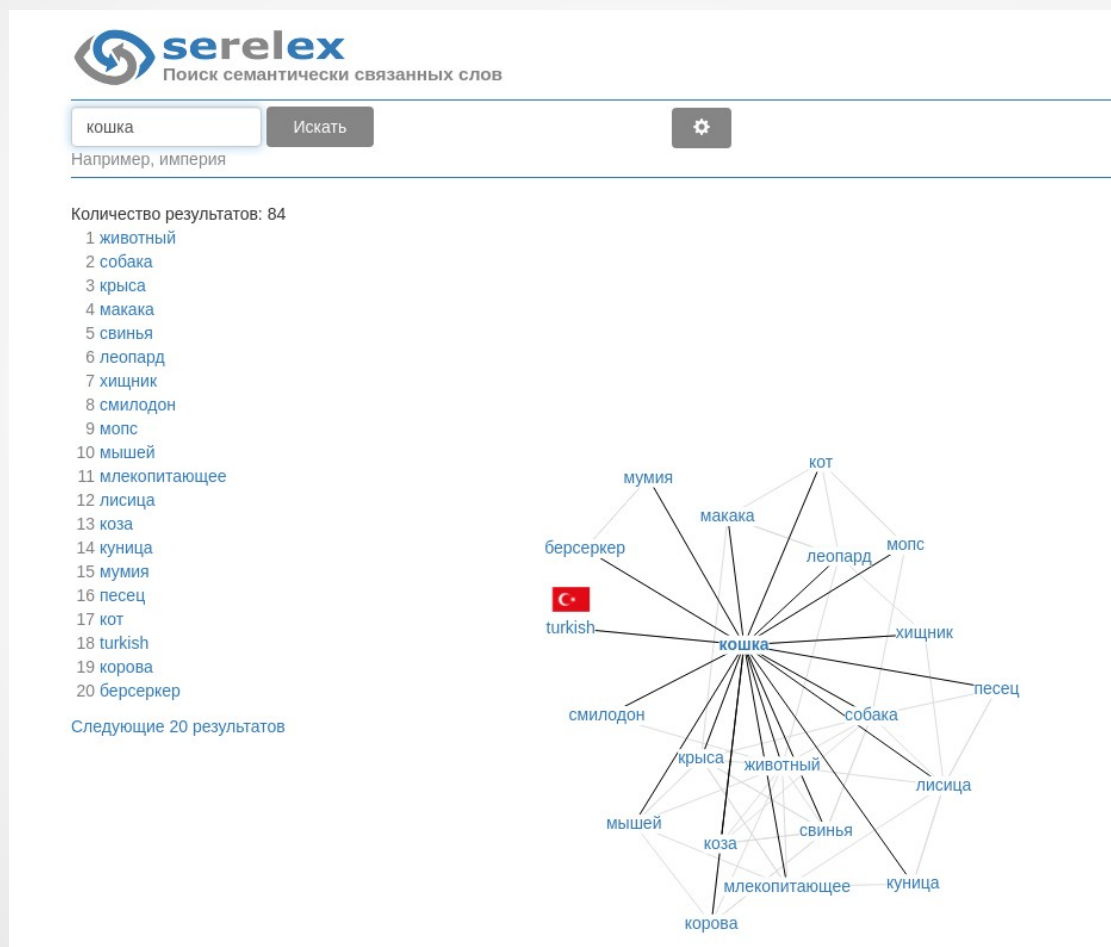
- <http://wordnet.princeton.edu/>
- <http://wordnet.ru/>

# Иерархии WordNet

```
Sense 3
bass, basso --
(an adult male singer with the lowest voice)
=> singer, vocalist, vocalizer, vocaliser
  => musician, instrumentalist, player
    => performer, performing artist
      => entertainer
        => person, individual, someone...
          => organism, being
            => living thing, animate thing,
              => whole, unit
                => object, physical object
                  => physical entity
                    => entity
              => causal agent, cause, causal agency
                => physical entity
                  => entity
```

# Семантическая близость и разрешение неоднозначностей

## Serelex — тезаурус on-line



<http://www.serelex.org/>  
[https://nlp.ru/Russian\\_Distributional\\_Thesaurus](https://nlp.ru/Russian_Distributional_Thesaurus)



# Семантическая близость и разрешение неоднозначностей

## Разрешение лексической многозначности (РЛМ) Word sense disambiguation (WDS)

- выбрать одно из нескольких значений слова по его контексту
- можно свести к задаче классификации (ML)

## Разграничение значений слова Word sense discrimination

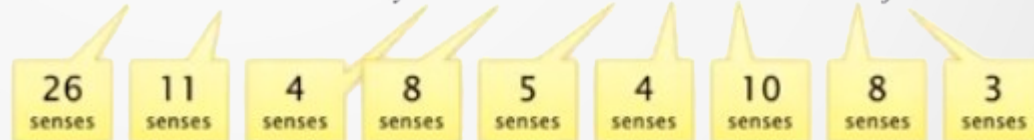
- по нескольким контекстам определить возможные значения слова
- можно свести к задаче кластеризации (ML)

# Семантическая близость и разрешение неоднозначностей

## Необходимо убирать неоднозначность

- озвучка текста
- информационный поиск
- автоматический перевод
- диалоговые системы

*I saw a man who is 98 years old and can still walk and tell jokes*



43,929,600  
senses

# Семантическая близость и разрешение неоднозначностей

## Разрешение лексической многозначности (РЛМ)

алгоритм Леска (1986) - РЛМ по словарю

берём все варианты определений искомого слова и слов его контекста

из всех вариантов [слово-значение] выбираем то, которое имеет наибольшее пересечение с определениями контекста

Пример: pine cone (сосновая шишка)

- *pine*
  1. a kind of **evergreen tree** with needle-shaped leaves
  2. to waste away through sorrow or illness
- *cone*
  1. A solid body which narrows to a point
  2. Something of this shape, whether solid or hollow
  3. Fruit of certain **evergreen trees**

# Семантическая близость и разрешение неоднозначностей

## Разрешение лексической многозначности (РЛМ)

классификатор контекста

каждого слова строим отдельный классификатор

признаки – слова контекста, их позиция и морфология

[ контекст ] → номер значения для слова в тезаурусе

проблема: слов очень много, есть редко употребляемые слова

# Семантическая близость и разрешение неоднозначностей

## Разрешение лексической многозначности (РЛМ)

оценка близости слова и контекста

заменяем задачу классификации каждого слова  
на задачу оценки близости слова и контекста

[ <контекст>, слово ] → оценка близости

оценка качества реализации метода  
производится по заранее размеченным данным

SENSEVAL – соревнование систем РЛМ

# Семантическая близость и разрешение неоднозначностей

Семантическая близость (similarity)

- автомобиль / мотоцикл

Семантическая связность (relatedness)

- автомобиль / бензин

будем употреблять термин «близость» для всех случаев

# Семантическая близость и разрешение неоднозначностей

## Оценка семантической близости

- использование тезауруса
- статистические модели (PMI)
- модели Word Embeddings

# Семантическая близость и разрешение неоднозначностей

Оценка семантической близости по тезаурусу

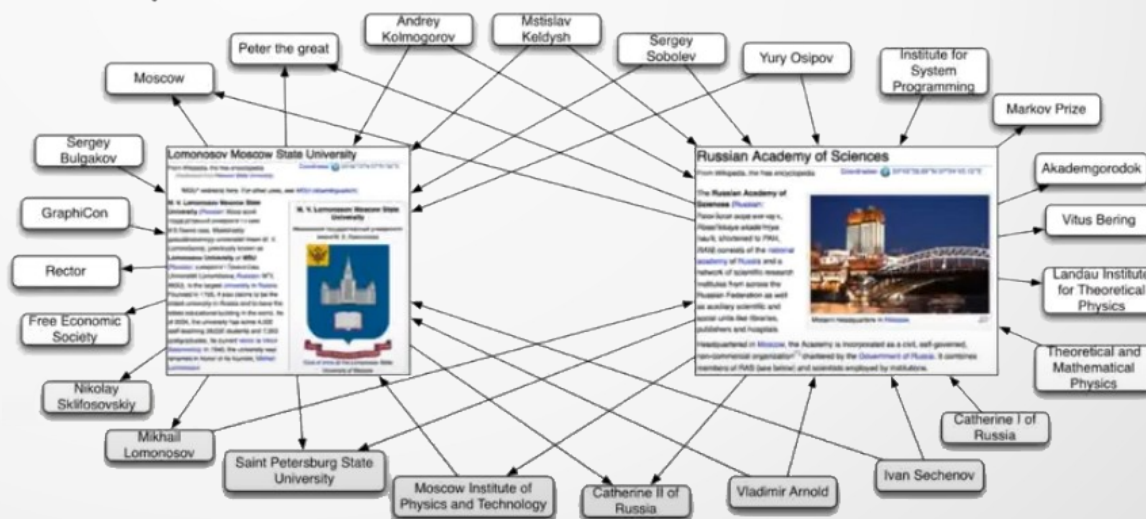
оцениваем расстояние по иерархии

метод Резника (1995)

метод Лина (1998)

## Использование Википедии

- Нормализованное количество общих соседей



- Близкие концепты чаще встречаются вместе



# Семантическая близость и разрешение неоднозначностей

## Статистическая оценка семантической близости

### Pointwise Mutual Information (PMI)

оценка совместного использования слов  $u$   $v$

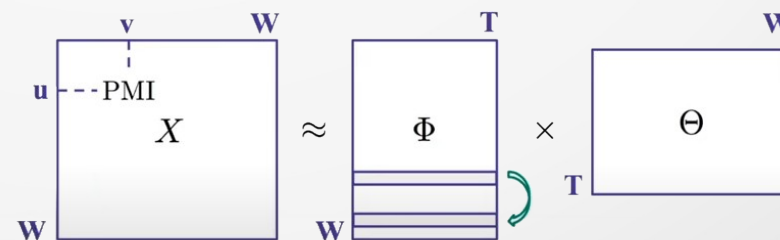
$$PMI(u, v) = \log \left( \frac{p(u, v)}{p(v)p(u)} \right)$$

$p(u, v)$  – частота использования словосочетания

$p(u)$  и  $p(v)$  - частота использования слов

	arts	boil	data	function	large	sugar	summarized	water
apricot	0	1	0	0	1	1	0	1
pineapple	0	1	0	0	1	1	0	1
digital	0	0	1	1	1	0	1	0
information	0	0	1	1	1	0	1	0

квадратная матрица контекстов



применим матричное разложение к квадратной матрице PMI  
для кодирования слов используем матрицу  $\Phi$

# Семантическая близость и разрешение неоднозначностей

Оценка семантической близости в семантических пространствах

Word Embeddings - кодирование слова по контексту

Word2Vec - совместно употребляемые в тексте слова  
отображаются в близкие точки пространства

$$w2v[\text{king}] - w2v[\text{man}] + w2v[\text{woman}] \approx w2v[\text{queen}]$$

Gensim — реализация на Python

построим ML-модель и обучим её кодировать слова по контексту

## Семантическая близость и разрешение неоднозначностей

подготовка данных Word2Vec – учитываем контекст слов.

- из текста  $T$  собираем словарь  $W$
- для каждого слова  $w$  собираем контекст (окрестность)  
т.е. слова удалённые от  $w$  не более чем на  $s$  позиций в  $T$
- выполняем унитарное кодирование(one-hot encoding)  $W$

$P_i:$   
0 0 1 0 0

$Q_i:$   
0 1 0 0 0  
0 0 0 0 1  
0 0 0 1 0  
1 0 0 0 0  
0 0 0 1 0

# NLP word2vec

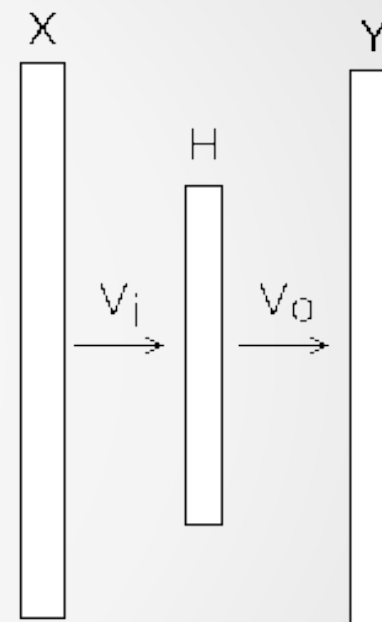
## нейросеть Word2Vec

размер входного слоя  $X$  = размеру словаря  $W$   
= размеру выходного слоя  $Y$

скрытый слой  $H$  - линейная активация

выходной слой  $Y$  - активация softmax

$$Y = \text{softmax}( (X \cdot V_i) \cdot V_o )$$



конечный результат - матрица  
внутренних представлений  $V_i$

# NLP word2vec

обучение сети word2vec

метод градиентного спуска

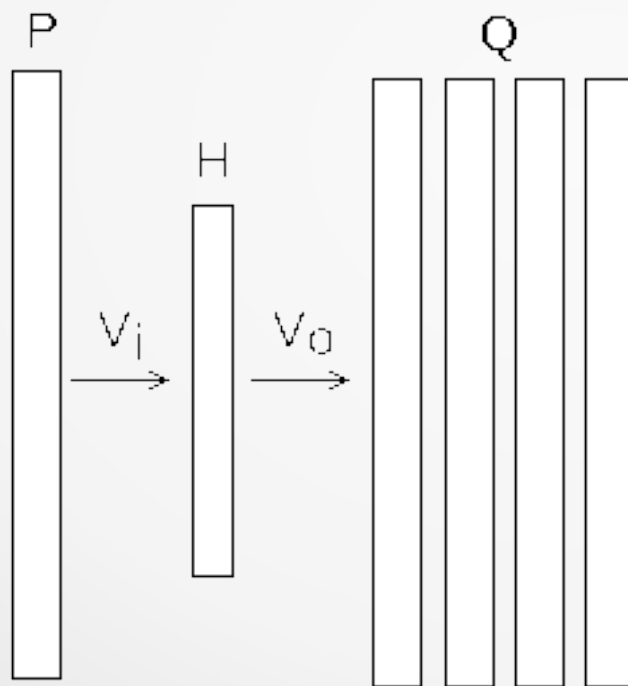
одна из двух стратегии

- Skip-Gram - по слову восстанавливаем контекст.
- CBOW(Continuous Bag of Words) по контексту восстанавливаем слово

# NLP word2vec

## обучение сети word2vec

- Skip-Gram - по слову восстанавливаем контекст.



# NLP word2vec

обучение сети word2vec - Skip-Gram - по слову восстанавливаем контекст.

1. на вход сети подаётся код слова  $P$ ,  
вычисляем состояние скрытого слоя  $H$   
вычисляем выход сети  $O$

2. вычисляем значение функции потерь

если значение потерь увеличилось  
то конец работы

$$E_i = \left| \log \sum \exp(U_i) - \sum_j \sum (U_i * Q_{ij}) \right|$$

3. для каждого слова контекста  $Q_j$  и входа  $P$ :

вычисляем ошибку  $D$  на выходе сети  $O$   
и изменение весов сети  $\Delta V_o$ ,  $\Delta V_i$

$$\begin{aligned} D &= O - Q_j \\ \Delta V_{oj} &= H^T \cdot D \\ \Delta V_{ij} &= D^T \cdot P \cdot V_o^T \end{aligned}$$

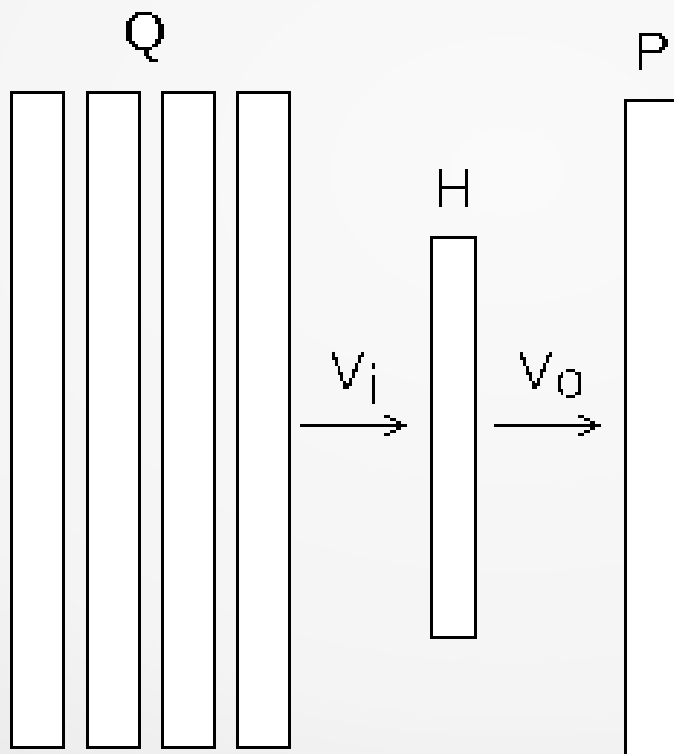
4. вычисляем суммарное изменение  
весов сети  $\Delta V_o$ ,  $\Delta V_i$   
корректируем веса  
и повторяем цикл для другого слова  $P$

$$\begin{aligned} \Delta V_o &= \sum_j \Delta V_{oj} \\ \Delta V_i &= \sum_j \Delta V_{ij} \end{aligned}$$

# NLP word2vec

## обучение сети word2vec

- CBOW(Continuous Bag of Words) по контексту восстанавливаем слово





# NLP word2vec

обучение сети word2vec - CBOW, по контексту восстанавливаем слово

1. на вход сети подаётся усреднённое значение контекста  $Q$ ,  
вычисляем состояние скрытого слоя  $H$   
вычисляем выход сети  $O$

$$H = \frac{1}{c} \sum_{j=1}^c Q_j \cdot V_i$$

$$U = H \cdot V_o$$
$$O = \text{softmax}(U)$$

2. вычисляем значение функции потерь

если значение потерь увеличилось  
то конец работы

$$E_i = \left| \log \sum \exp(U_i) - \sum (U_i * P_i) \right|$$

3. для каждого слова контекста  $Q_j$  и кода слова  $P$ ,  
вычисляем ошибку  $D$  на выходе сети  $O$   
и изменение весов сети  $\Delta V_o$ ,  $\Delta V_i$ .

$$D = O - P$$
$$\Delta V_o = H^T \cdot D$$
$$\Delta V_i = \sum_j D^T \cdot Q_j \cdot V_o^T$$

4. корректируем веса  
и повторяем цикл для другого слова  $P$

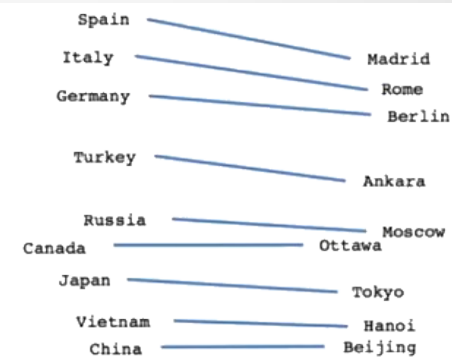
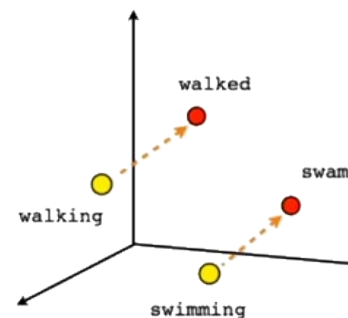
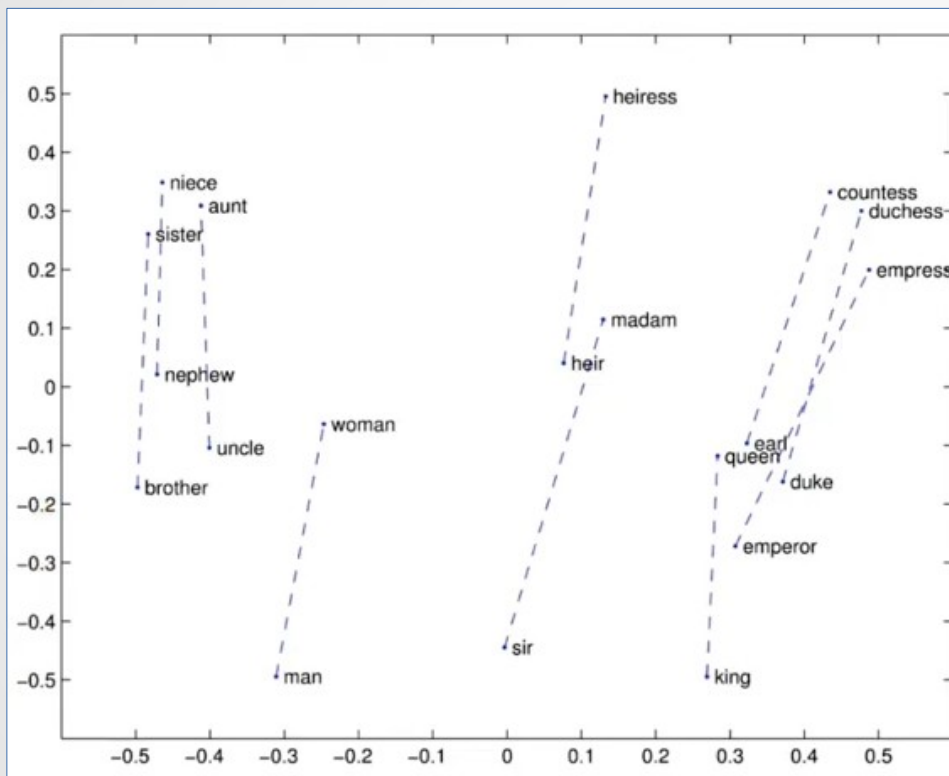
# NLP word2vec

## Результат работы

```
{ 'состоится': ['конечно', 'бал'],  
  'волнения': ['описывать', 'виргинский'],  
  'печатных': ['листа', 'липутину'],  
  'лямшин': ['вечер', 'чуткости'],  
  'или': ['отчасти', 'нет'],  
  'крышу': ['под', 'надлежащего'],  
  'общее': ['пригорюнясь', 'пострадать'],  
  'степанович': ['осилил', 'новых'],  
  'часов': ['кучке', 'часу'],  
  'вас': ['для', 'завтра'],  
  'слишком': ['милости', 'была'],  
  'крикнул': ['попа', 'прочтет'],  
  'видеть': ['свое', 'будто'],  
  'наготове': ['кармазинове', 'беспокойством'],  
  'помещался': ['этою', 'подыматься'],  
  'выполнено': ['планете', 'несмотря'],  
  'тебя': ['люблю', 'взрывов'],  
  'началом': ['фунтов', 'центру'],  
  'страдал': ['заявите', 'тесно'],  
  'отчаянная': ['новая', 'кончив']}
```

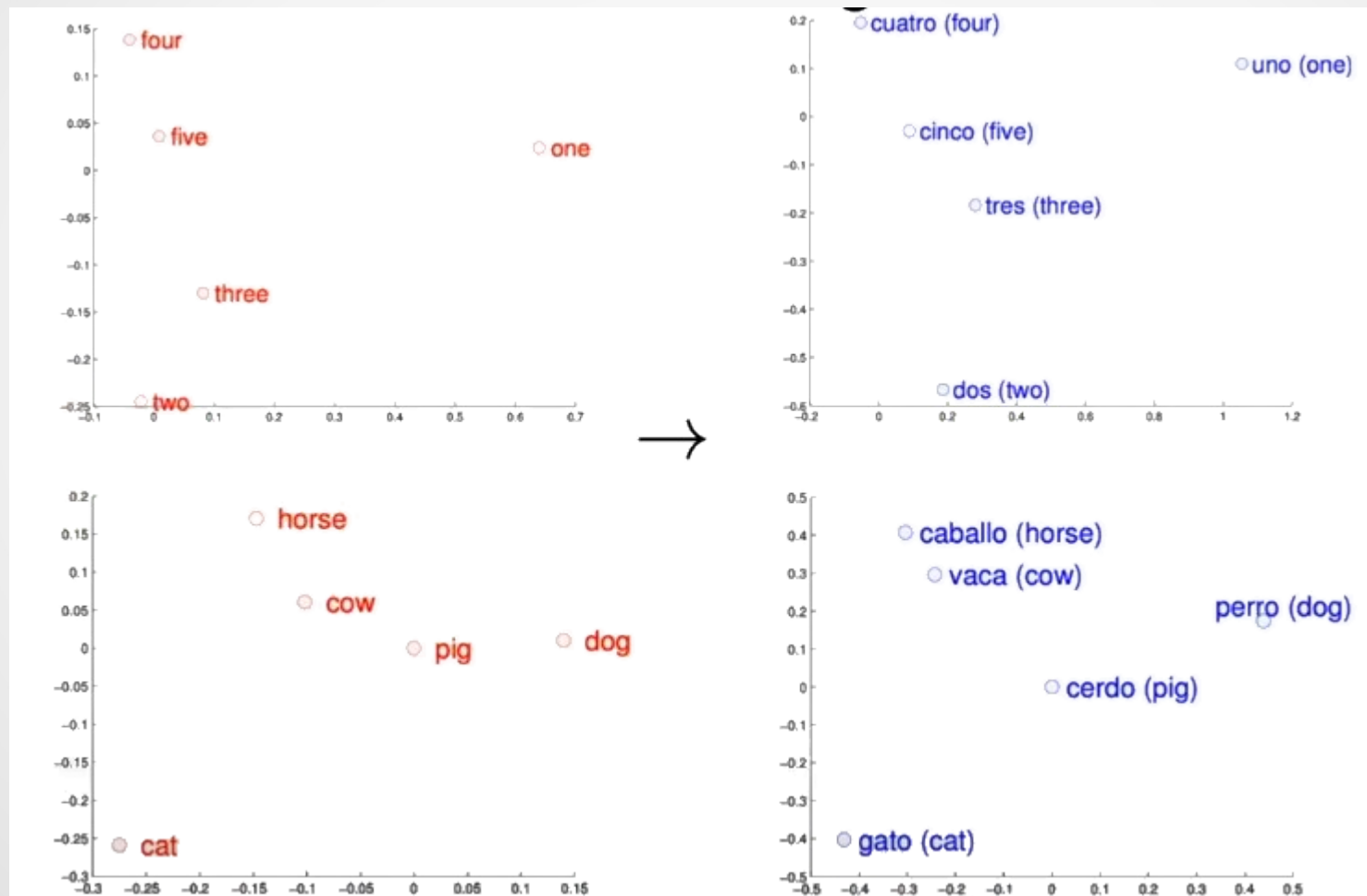
# Семантическая близость и разрешение неоднозначностей

близкие по контексту слова отображаются в близкие точки w2v



# Семантическая близость и разрешение неоднозначностей

взаимное расположение w2v в разных языках схожи

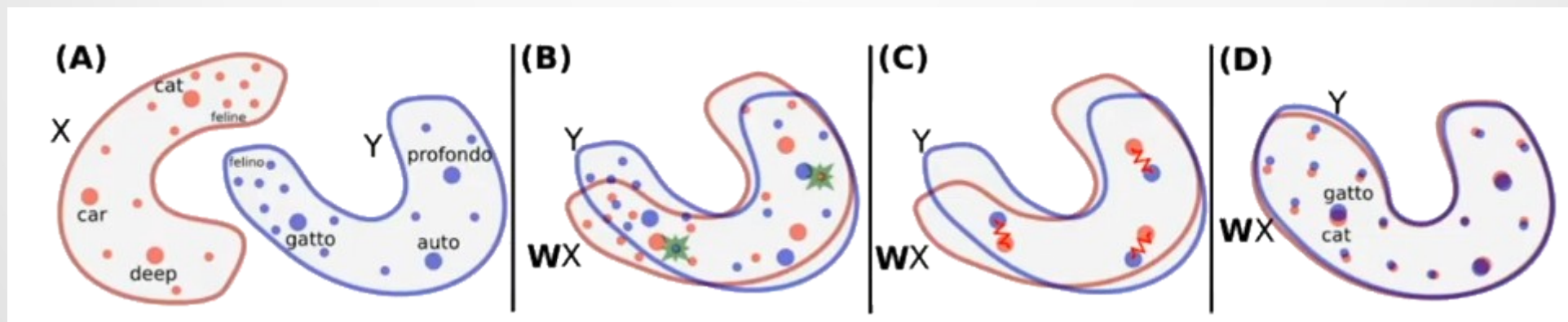


# Семантическая близость и разрешение неоднозначностей

взаимное расположения  $w_2v$  в разных языках схожи

зная перевод некоторых слов и на основе этого построив отображение из  $w_2v$  пространства одного языка в другое,

мы получаем перевод всех остальных слов на основе контекста



# Семантическая близость и разрешение неоднозначностей

## Литература

git clone [https://github.com/mechanoid5/ml\\_nlp.git](https://github.com/mechanoid5/ml_nlp.git)

Турдаков Д.Ю.

Основы обработки текстов. лекция 9. Лексическая семантика. ИСП РАН, 2017

<https://www.youtube.com/watch?v=IaIgSdJD5nE>

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean  
Distributed Representations of Words and Phrases and their Compositionality

Радослав Нейчев Прикладное машинное обучение 1.Intro to NLP. Word embeddings -  
Лекторий ФПМИ

[https://www.youtube.com/watch?v=aZ5se\\_SW81c](https://www.youtube.com/watch?v=aZ5se_SW81c)

Евгений Борисов О методе кодирования слов word2vec.

<http://mechanoid.su/ml-w2v.html>

Kuzma Khrabrov

Применение сиамских нейросетей в поиске.

<https://habr.com/ru/company/mailru/blog/468075/>