Семантическая близость и модели Word Embeddings

Евгений Борисов

NLP Word Embeddings

Уровни сложности при автоматической обработке текстов

Прагматика (Дискурс) - смысловые контексты
Семантика - смыслы последовательностей слов
Синтаксис - правила формирования последовательностей слов
Морфология - отдельные слова и устойчивые словосочетания

NLP Word Embeddings

Семантика

- лексическая, отдельные слова

- композиционная, комбинации слов

задачи

- разрешение многозначности

- оценка семантической близости

NLP Word Embeddings

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.

Отношения между словами

синонимия - общий смысл

машина, автомобиль

антонимия - противоположность

большой / маленький, вверх / вниз

гипонимия - обобщение

яблоко / фрукт, овчарка / собака

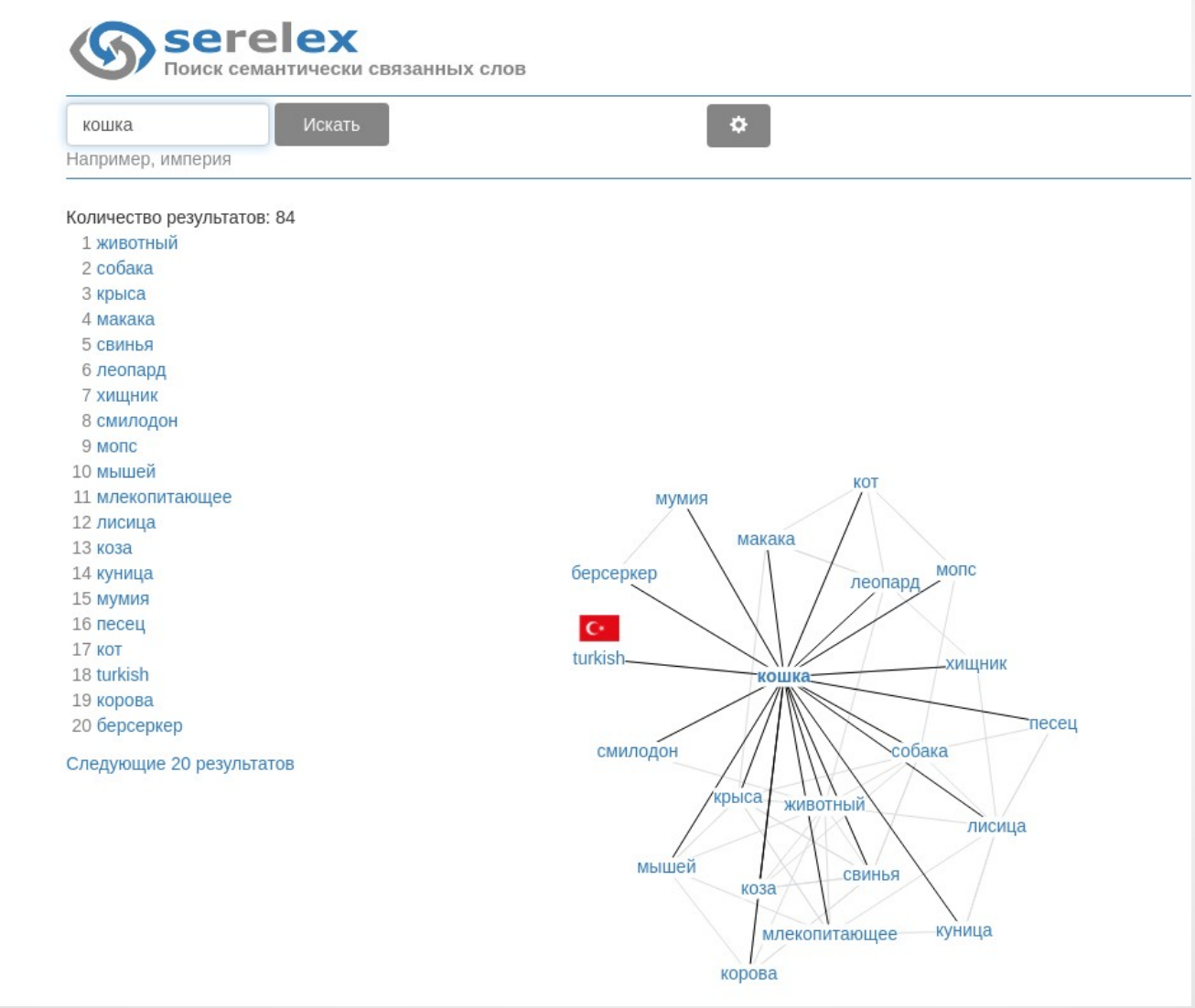
партономия - часть, вхождение

колесо / автомобиль, житель / город

NLP Word Embeddings

Тезаурус — словарь со связями

http://www.serelex.org/
https://nlpub.ru/Russian_Distributional_Thesaurus



NLP Word Embeddings

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

www.youtube.com/watch?v=IaIgSdJD5nE

NLP Word Embeddings

Семантическая близость (similarity)

- автомобиль / мотоцикл

Семантическая связность (relatedness)

- автомобиль / бензин

будем употреблять термин «**близость**» для всех случаев

Оценка семантической близости

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

NLP Word Embeddings

Оценка семантической близости по тезаурусу

оцениваем расстояние по иерархии

метод Резника (1995)

метод Лина (1998)

NLP Word Embeddings

Статистическая оценка семантической близости

	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

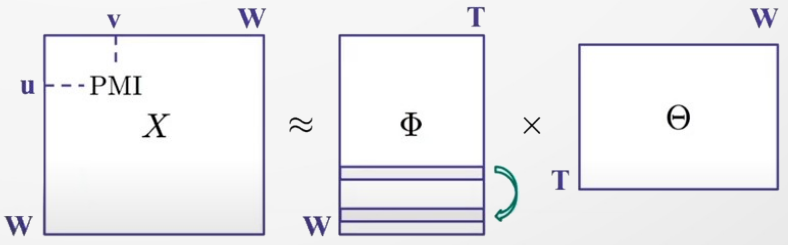
Pointwise Mutual Information (PMI)

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



применим матричное разложение к квадратной матрице PMI
для кодирования слов используем матрицу Φ

NLP Word Embeddings

Оценка близости в семантических пространствах

Word Embeddings - кодирование слова по контексту

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

w2v[king] – w2v[man] + w2v[woman] ≈ w2v[queen]

Gensim – реализация на Python

построим ML-модель и обучим её кодировать слова по контексту

NLP Word Embeddings

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

Pi:
0 0 1 0 0

Qi:
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 Word Embeddings

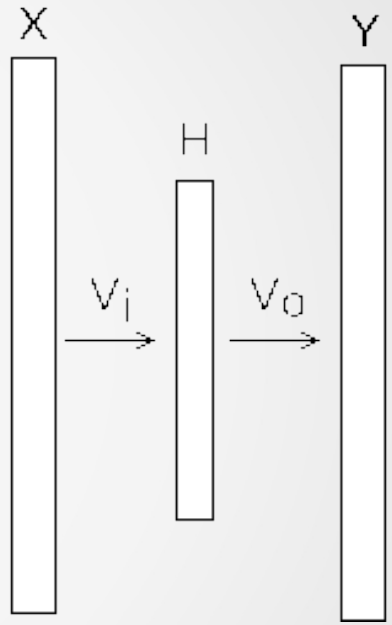
нейросеть Word2Vec

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

скрытый слой H - линейная активация

выходной слой Y – активация softmax

$$Y=softmax((X \cdot Vi) \cdot Vo)$$



конечный результат - матрица внутренних представлений *Vi*

NLP Word Embeddings

обучение сети word2vec

метод градиентного спуска

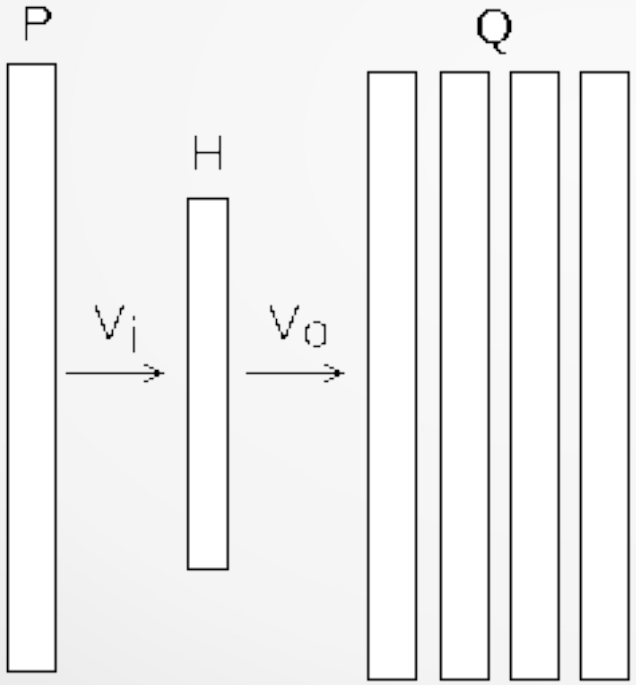
одна из двух стратегии

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

NLP Word Embeddings

обучение сети word2vec

- Skip-Gram - по слову восстанавливаем контекст.



NLP Word Embeddings

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 потери

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

если значение потери увеличилость
то конец работы

3. для каждого слова контекста Q_j и входа P:

вычисляем ошибку D на выходе сети O
и изменение весов сети ΔV_o, ΔV_i

$$D = O - Q_j$$
$$\Delta Vo_j = H^T \cdot D$$
$$\Delta Vi_j = D^T \cdot P \cdot Vo^T$$

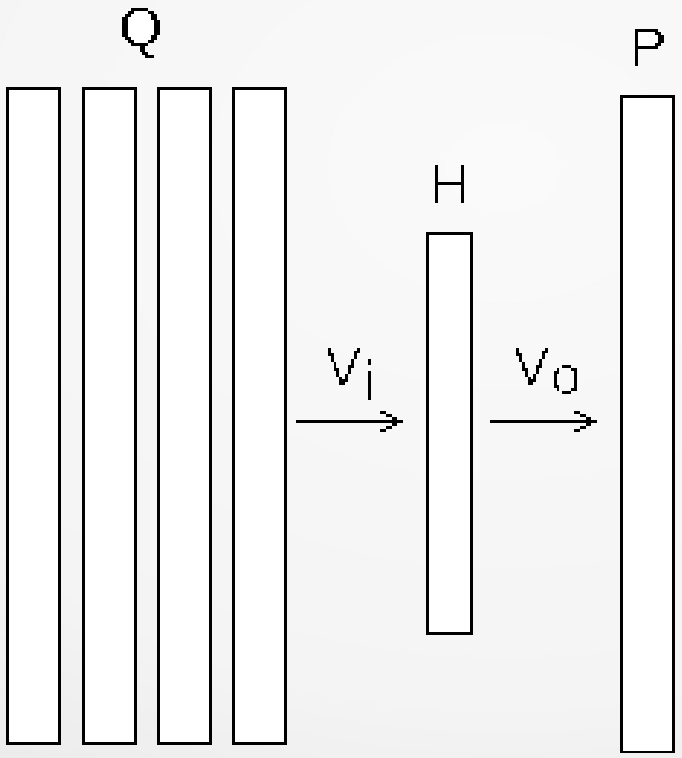
4. вычисляем суммарное изменение
весов сети ΔV_o, ΔV_i
корректируем веса
и повторяем цикл для другого слова P

$$\Delta Vo = \sum_j \Delta Vo_j$$
$$\Delta Vi = \sum_j \Delta Vi_j$$

NLP Word Embeddings

обучение сети word2vec

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



NLP Word Embeddings

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

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

$$U = H \cdot V_o$$
$$O = softmax(U)$$

2.вычисляем значение функции потери

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

если значение потери увеличилось
то конец работы

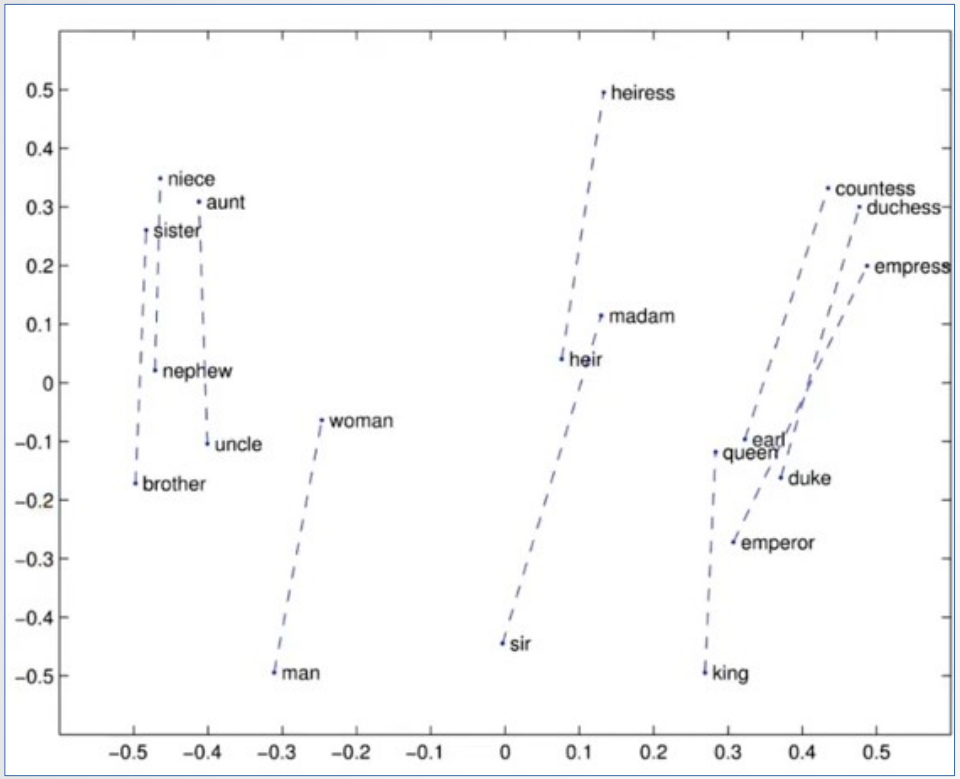
3. для каждого слова контекста Qj и кода слова P,
вычисляем ошибку D на выходе сети O
и изменение весов сети ΔVo, ΔVi.

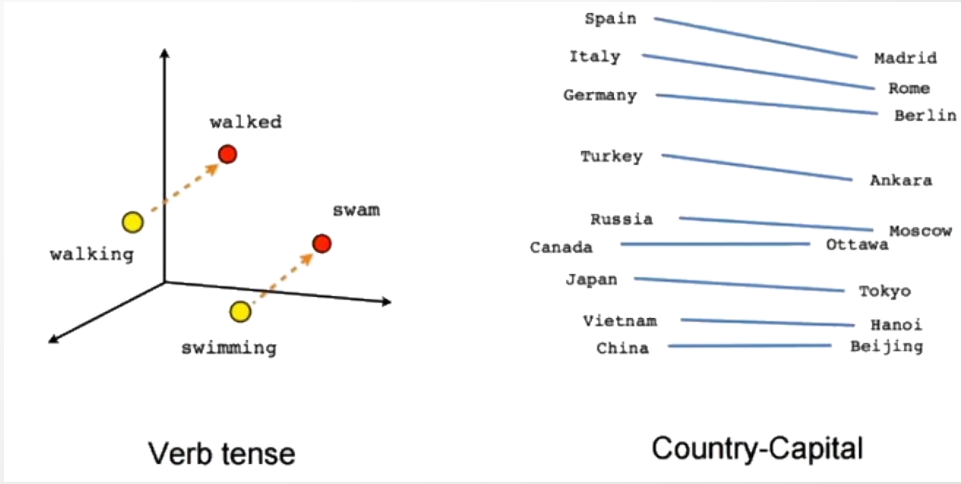
$$D = O - P$$
$$\Delta Vo = H^T \cdot D$$
$$\Delta Vi = \sum_j D^T \cdot Q_j \cdot Vo^T$$

4. корректируем веса
и повторяем цикл для другого слова P

NLP Word Embeddings

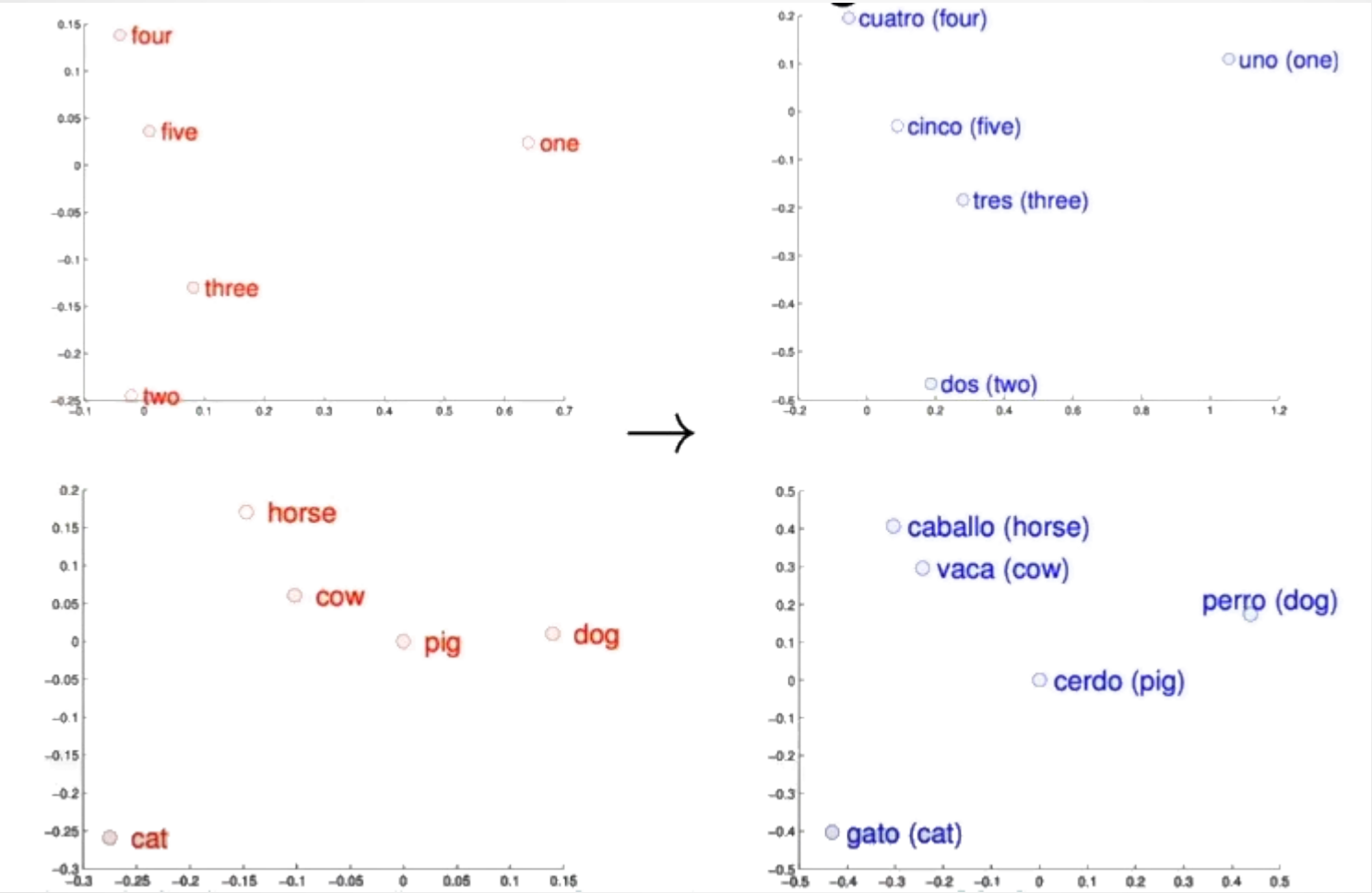
близкие по контексту слова отображаются в близкие точки w2v





NLP Word Embeddings

взаимное расположение w2v в разных языках схожи

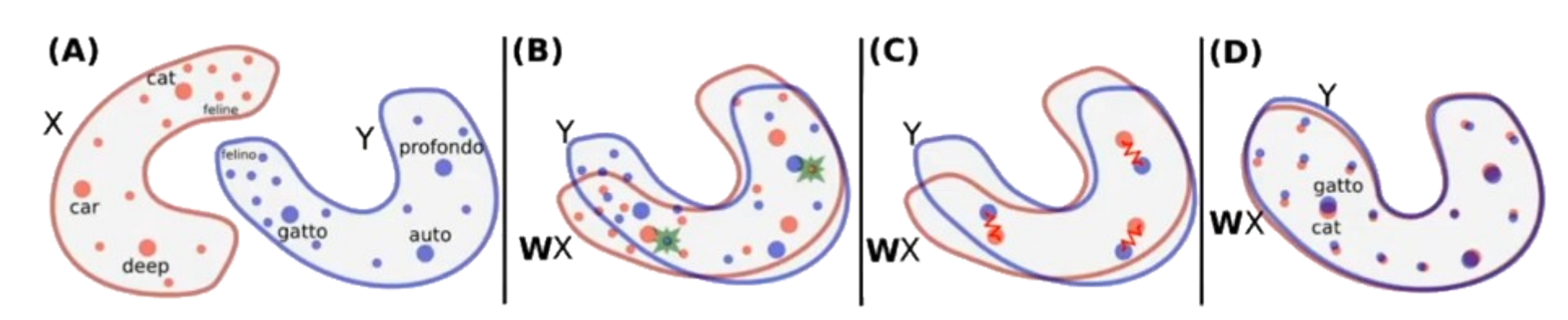


NLP Word Embeddings

взаимное расположения w2v в разных языках схожи

зная перевод некоторых слов и на основе этого построив отображение из w2v пространства одного языка в другое,

мы получаем перевод всех остальных слов на основе контекста



NLP Word Embeddings

Литература

git clone https://github.com/mechanoid5/ml_lectorium

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

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/