



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

Оценка семантической близости

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

Семантическая близость и разрешение неоднозначностей

Статистическая оценка семантической близости

Pointwise Mutual Information (PMI)

оценка совместного использования слов u и v

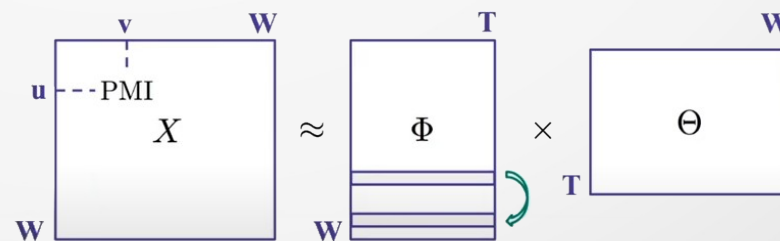
$$PMI(u, v) = \log \left(\frac{p(u, v)}{p(v)p(u)} \right)$$

$p(u, v)$ – 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 Φ

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

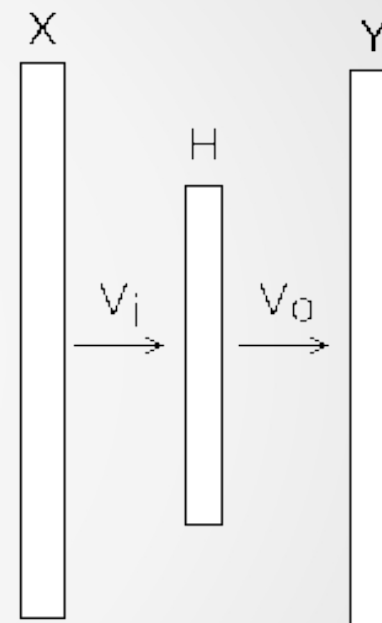
нейросеть Word2Vec

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

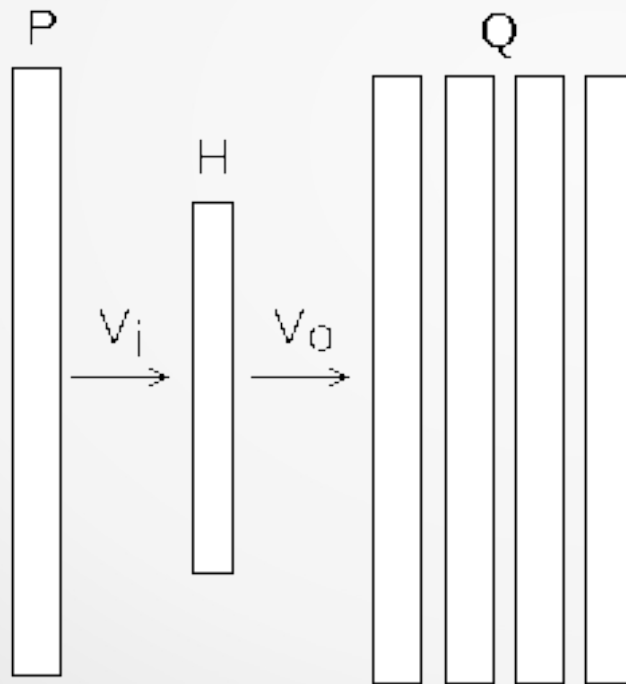
одна из двух стратегии

- 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

3. для каждого слова контекста Q_j и входа P :

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

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

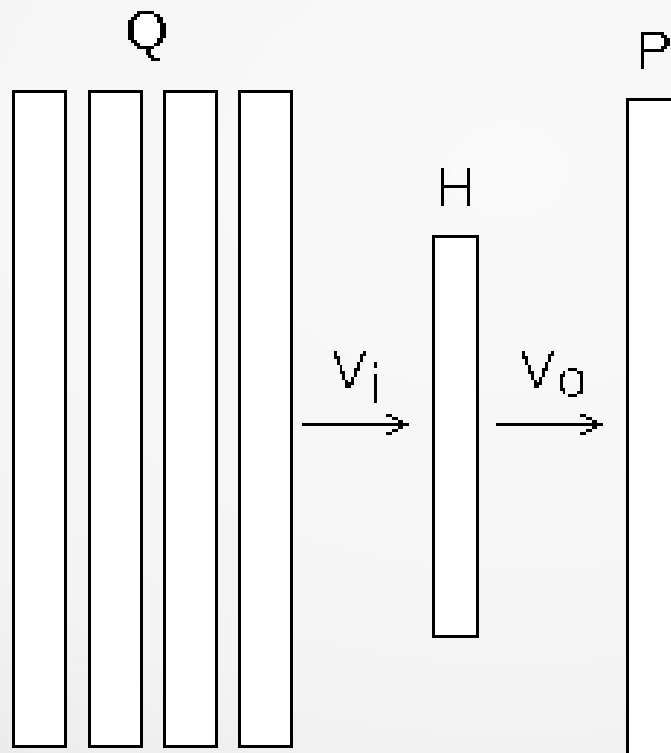
4. вычисляем суммарное изменение
весов сети $\Delta V_o, \Delta V_i$
корректируем веса
и повторяем цикл для другого слова P

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

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

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

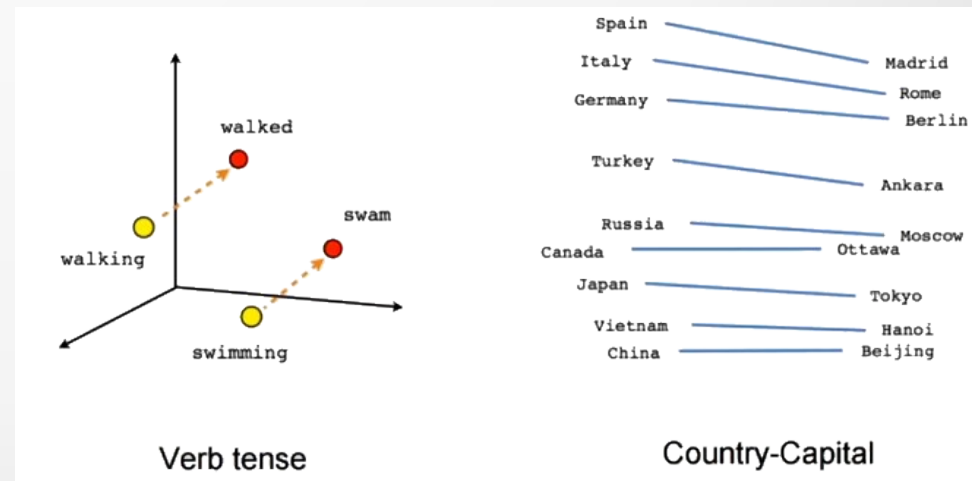
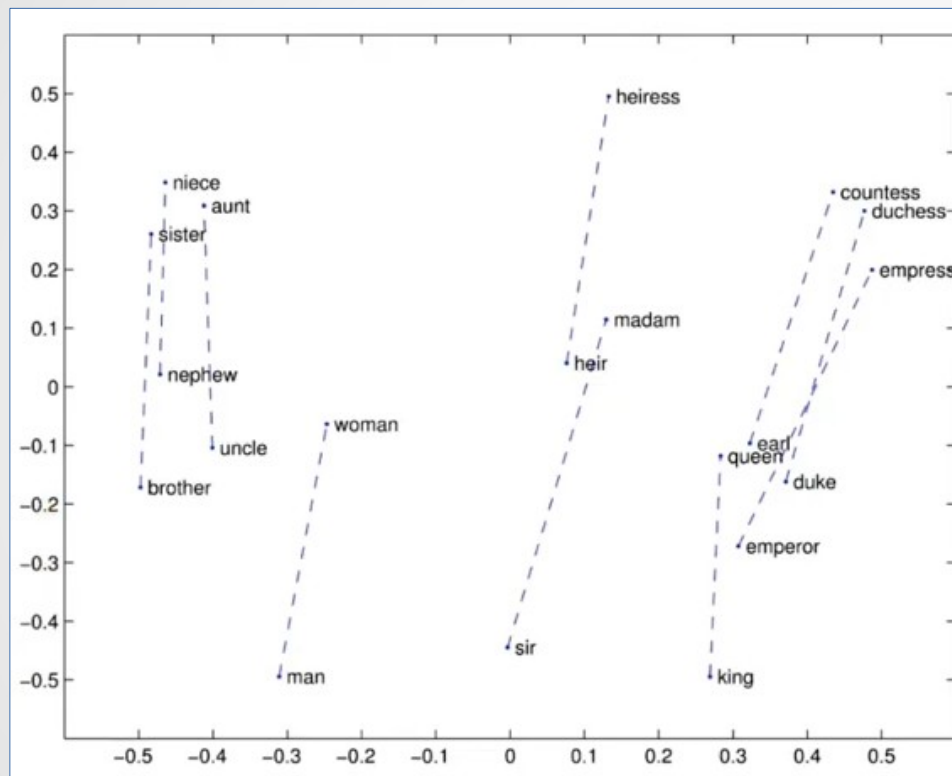
3. для каждого слова контекста Q_j и кода слова P ,
вычисляем ошибку D на выходе сети O
и изменение весов сети ΔV_o , ΔV_i .

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

4. корректируем веса
и повторяем цикл для другого слова P

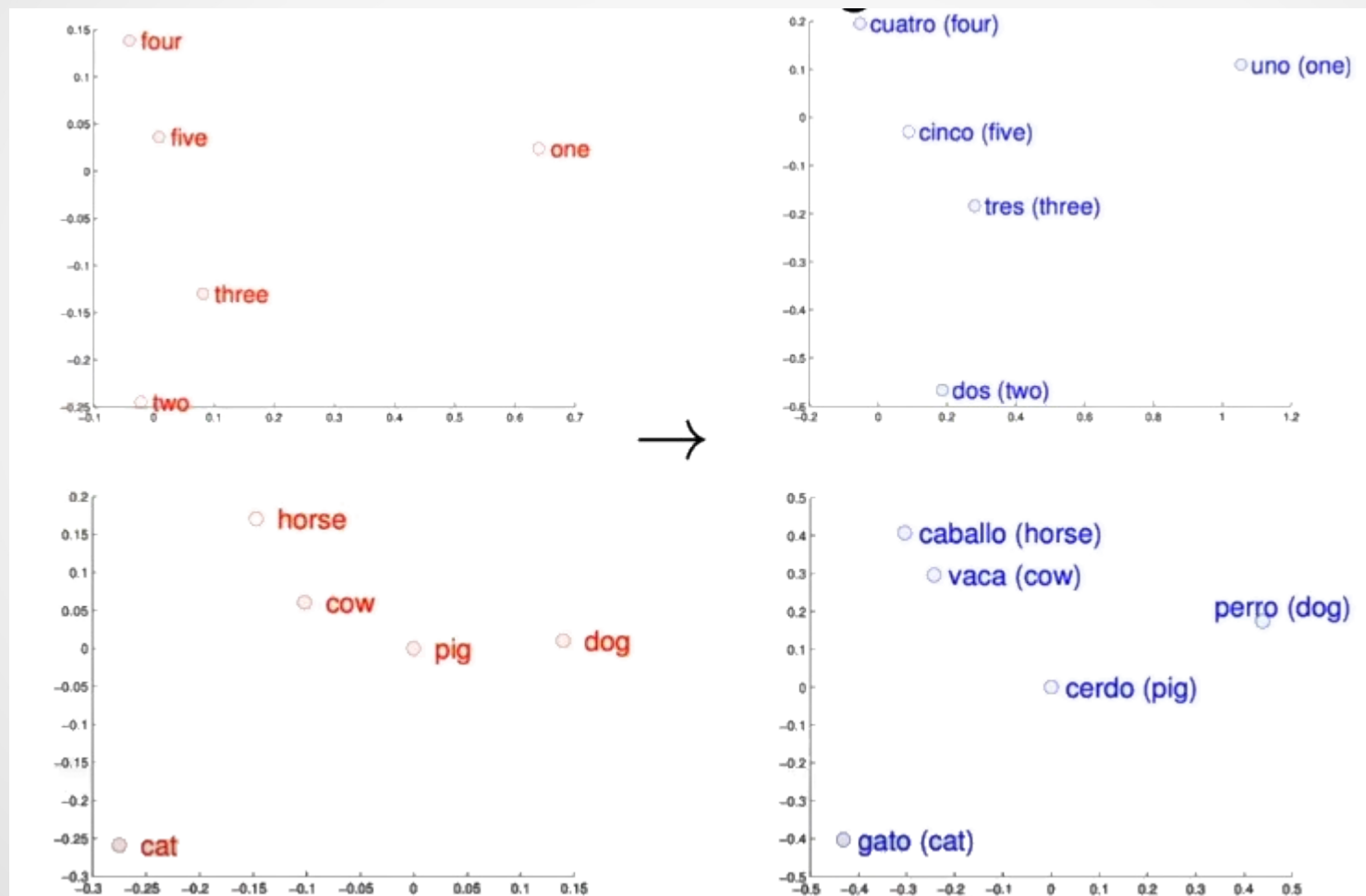
Семантическая близость и разрешение неоднозначностей

близкие по контексту слова отображаются в близкие точки $w2v$



Семантическая близость и разрешение неоднозначностей

взаимное расположение w2v в разных языках схожи

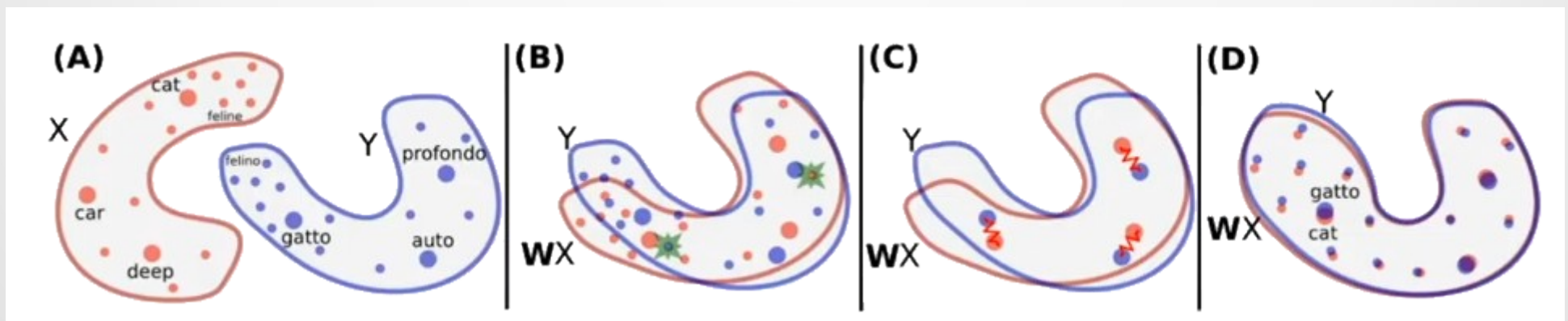


Семантическая близость и разрешение неоднозначностей

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

Борисов Е.С. Методы машинного обучения. 2024

https://github.com/mechanoid5/ml_lectorium_2024_I

Борисов Е.С. Методы обработки текстов на естественном языке. 2024

https://github.com/mechanoid5/ml_nlp_2024_I

Турдаков Д.Ю. Основы обработки текстов. 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

Jay Alammar Word2vec в картинках. (перевод - Анатолий Ализар)

<https://habr.com/ru/articles/446530/>

Евгений Борисов 0 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/>