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

www.youtube.com/watch?v=IaIgSdJD5nE

Семантическая близость и разрешение неоднозначностей

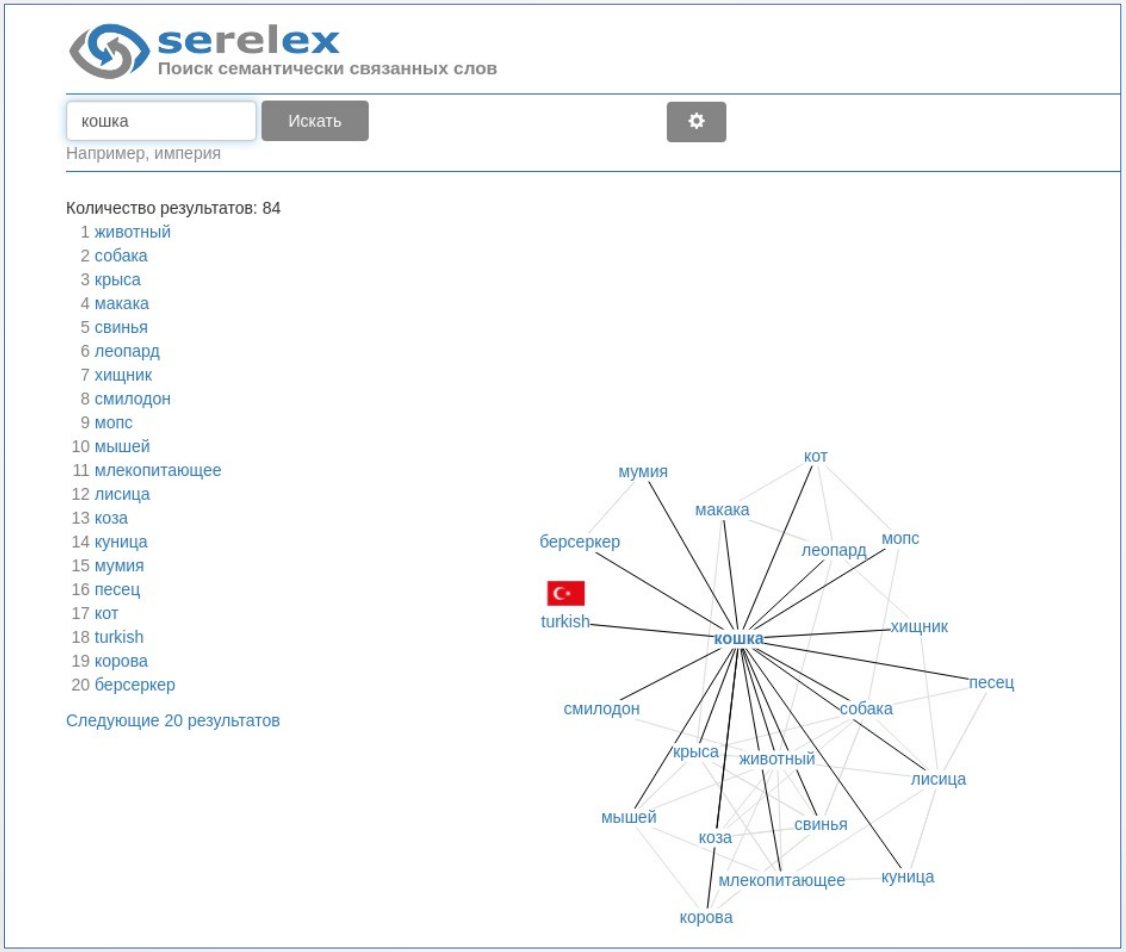
Иерархии WordNet

```
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
```

www.youtube.com/watch?v=IaIgSdJD5nE

Семантическая близость и разрешение неоднозначностей

Serelex — тезаурус on-line



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

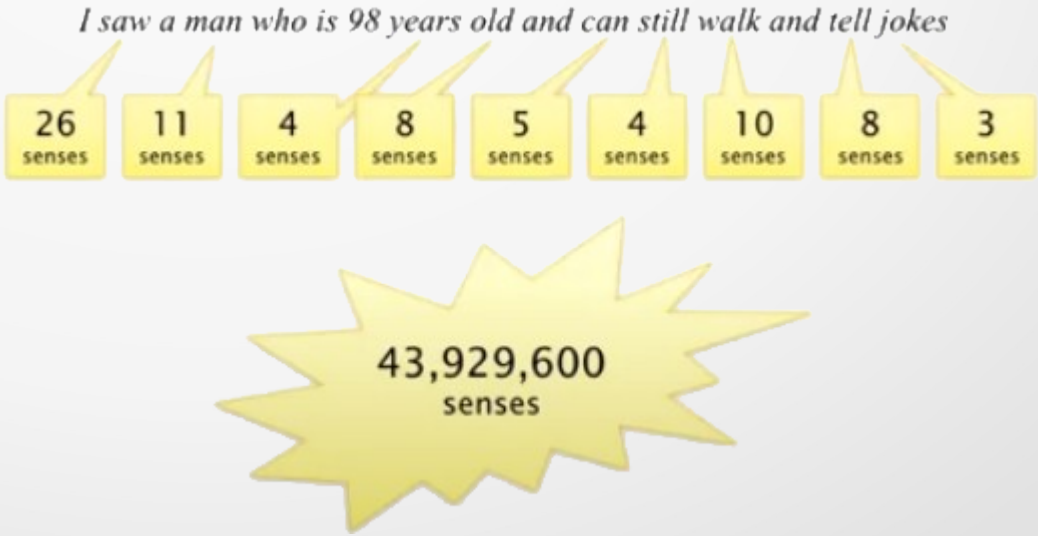
Необходимо убирать неоднозначность

- озвучка текста

- информационный поиск

- автоматический перевод

- диалоговые системы



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)

Семантическая близость и разрешение неоднозначностей

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

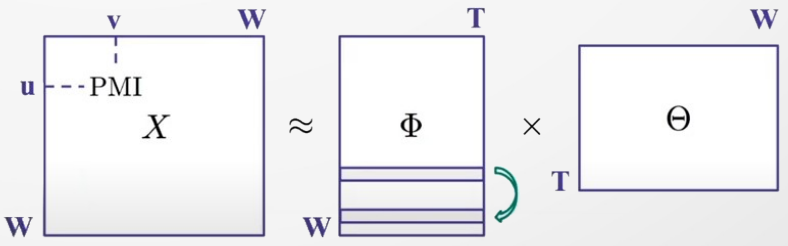
Pointwise Mutual Information (PMI)

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

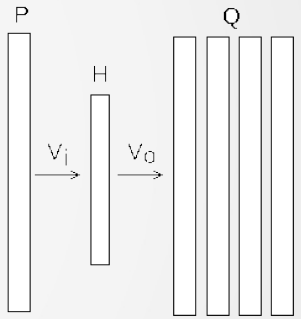
p(u) и *p(v)* - частота использования слов



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

Семантическая близость и разрешение неоднозначностей

Оценка семантической близости в семантических пространствах



Word Embeddings - кодирование слова по контексту

Word2Vec - совместно употребляемые в тексте слова из W отображаются в близкие точки пространства V

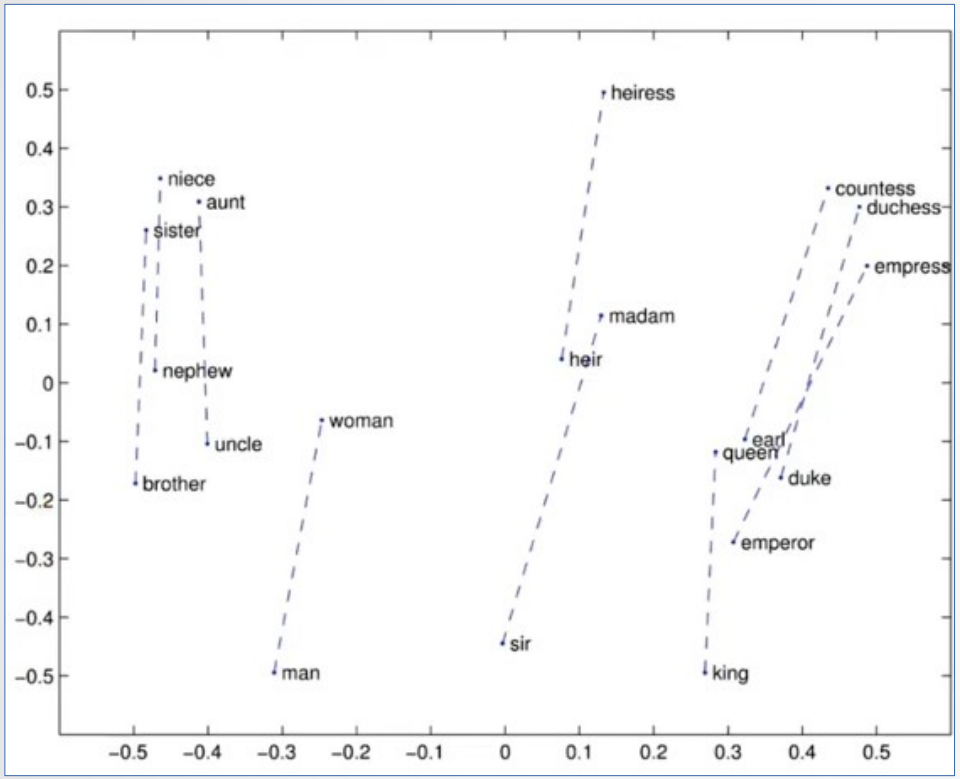
w2v[king] − w2v[man] + w2v[woman] ≈ w2v[queen]

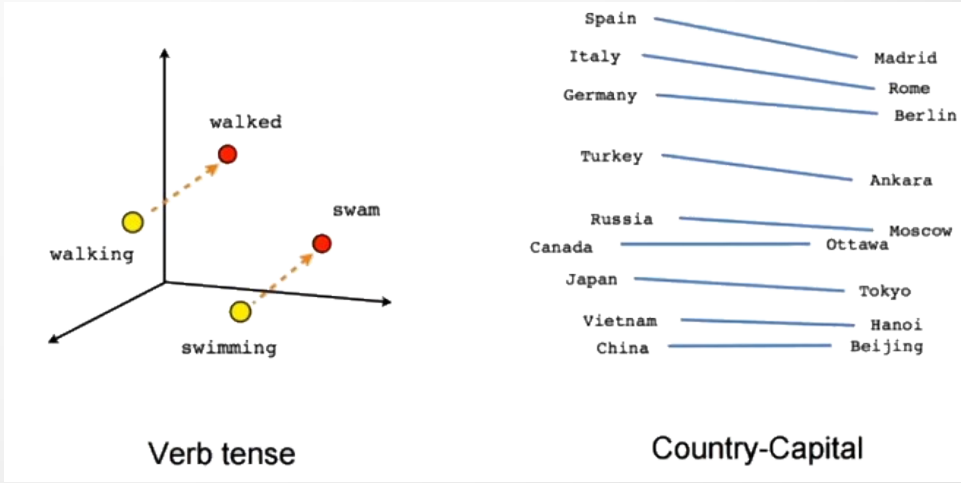
Word2Vec Skip-Gram – обучаем модель по слову восстанавливать контекст

Gensim – реализация на Python

Семантическая близость и разрешение неоднозначностей

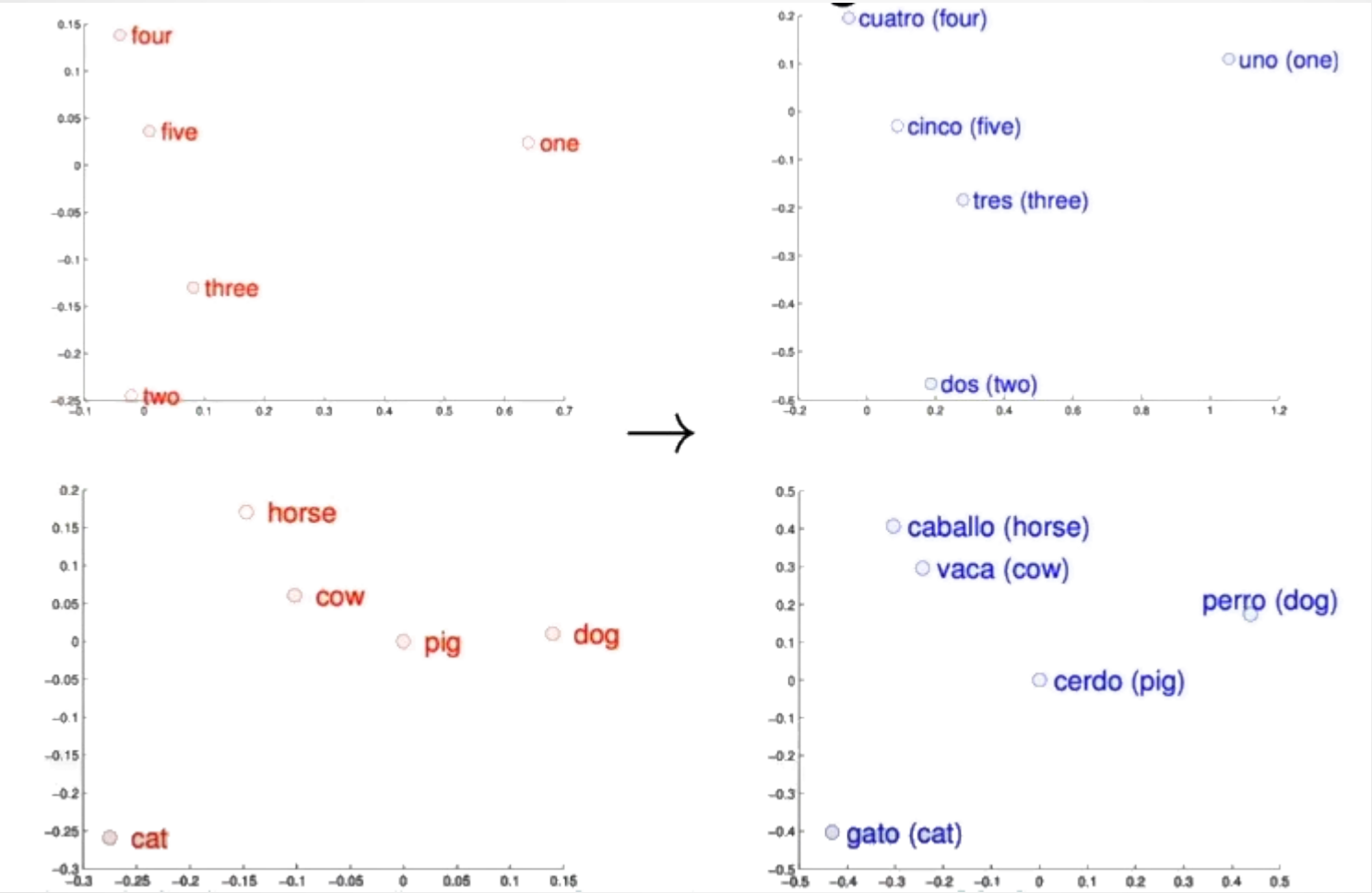
близкие по контексту слова отображаются в близкие точки w2v





Семантическая близость и разрешение неоднозначностей

взаимное расположение w2v в разных языках схожи

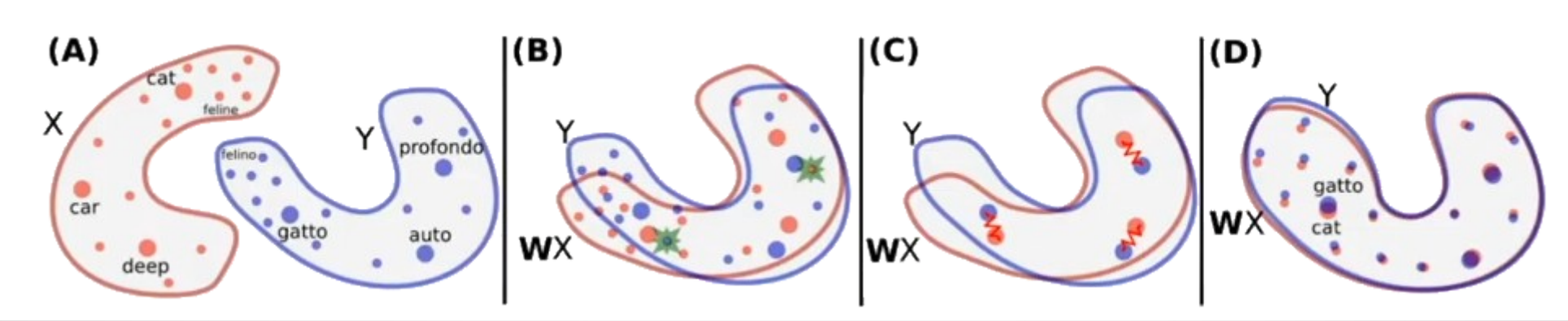


Семантическая близость и разрешение неоднозначностей

взаимное расположения w2v в разных языках схожи

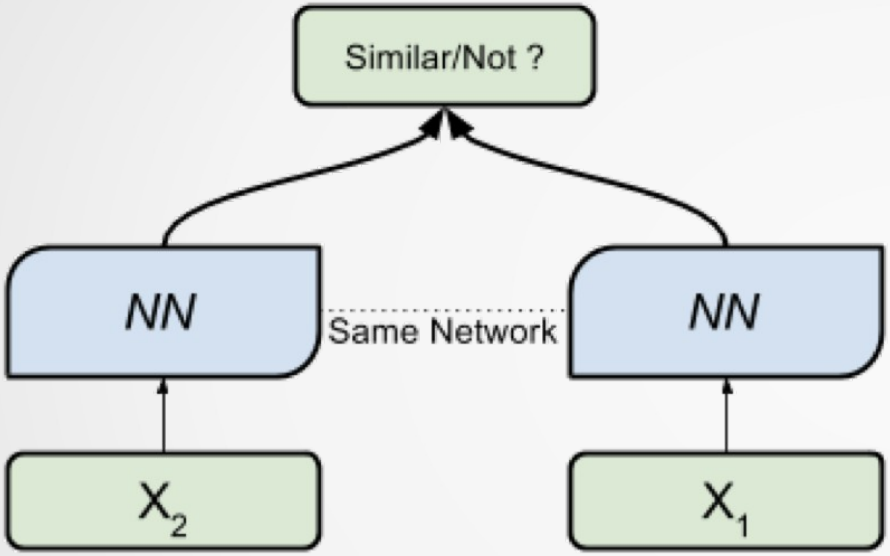
зная перевод некоторых слов и на основе этого построив отображение из w2v пространства одного языка в другое,

мы получаем перевод всех остальных слов на основе контекста



Семантическая близость и разрешение неоднозначностей

Siamese neural network / Сиамские нейросети



Metric learning

оценка схожести объектов

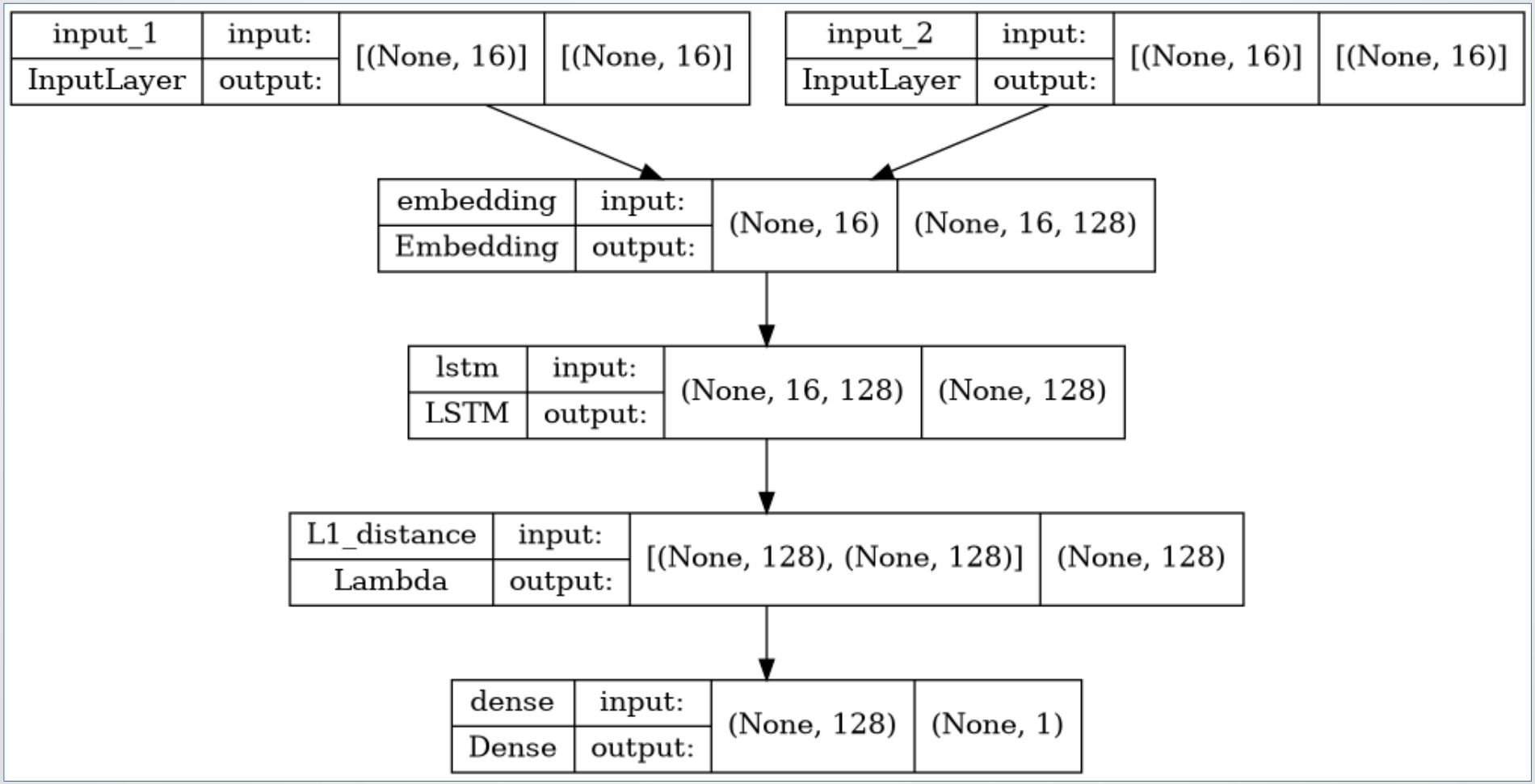
Contrastive loss - функция потери основанная на метрике D

$$L_{contrast} = (1 - Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\max(0, m - D_W)^2$$

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

Семантическая близость и разрешение неоднозначностей

Пример сиамской нейросети - определяем схожесть текстов



Семантическая близость и разрешение неоднозначностей

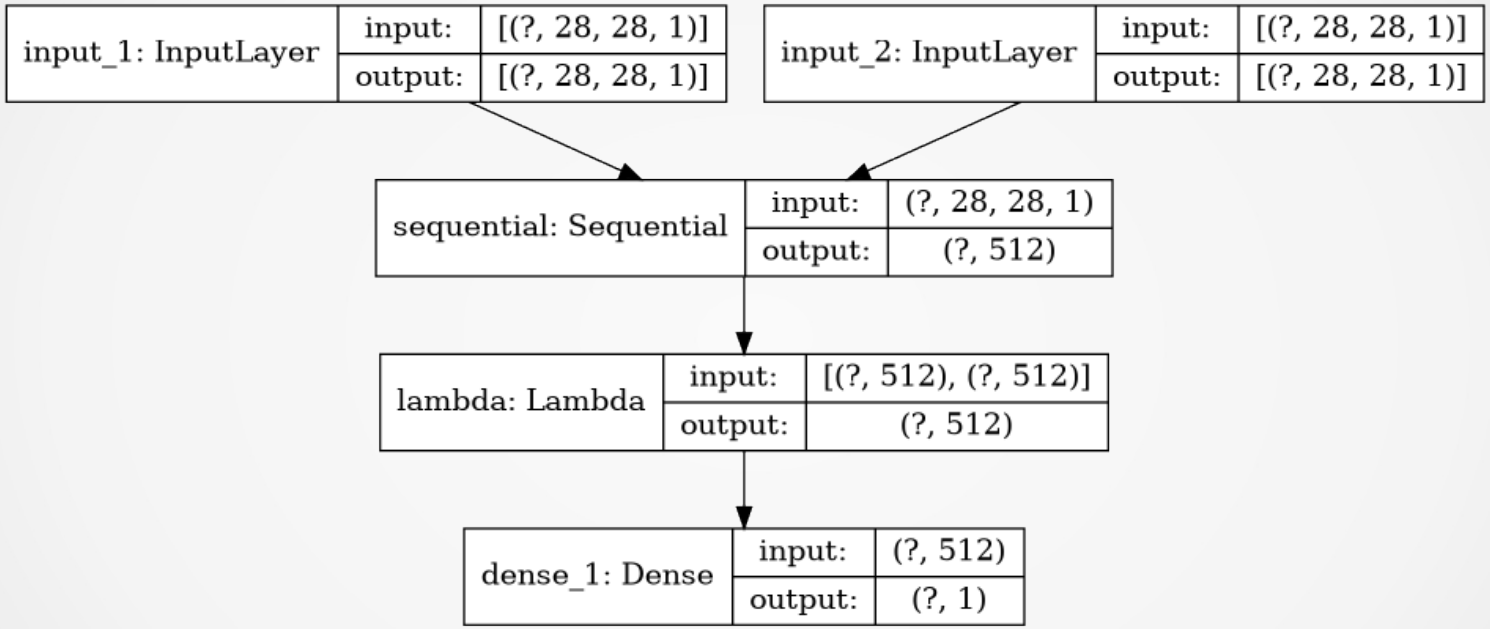
Пример сиамской нейросети - определяем схожесть текстов

ttext_L	ttext_R	dist
Ограды (заборы) и ограждения прочие	Сооружения для занятий спортом и отдыха	0.0
Услуги центрального аппарата Следственного комитета Российской Федерации	Услуги федеральных арбитражных судов округов	0.0
Работы по укладке ковровых покрытий, линолеума и прочих гибких материалов для покрытия полов	Работы по устройству полов из тераццо, работы с использованием мрамора, гранита и сланца	0.0
Дистилляты прочие полного цикла производства	Ликероводочные изделия крепостью свыше 25 % прочие	0.0

ttext_L	ttext_R	dist
Мотоциклы с поршневым двигателем внутреннего сгорания с рабочим объемом цилиндров не более 50 см3	Услуги по ремонту электрооборудования прочих автотранспортных средств	1.263635
Услуги по чистовой обработке прочих стеклянных изделий, включая технические стеклянные изделия	Услуги по заграничным и каботажным перевозкам морскими судами сухих сыпучих грузов	1.440093
Нефть смесевая особо высокосернистая особо легкая	Рыба и филе рыбное холодного копчения	1.388784
Услуги по производству ювелирных и соответствующих изделий отдельные, выполняемые субподрядчиком	Услуги по бронированию и взаимосвязанные услуги прочие	1.094095
Игры и игрушки, не включенные в другие группировки	Панели и прочие комплекты электрической аппаратуры коммутации или защиты на напряжение не более 1 кВ	1.318040
Услуги по сбору неопасных отходов городского хозяйства, непригодных для повторного использования	Трубы и муфты асбестоцементные безнапорные	1.437839

Семантическая близость и разрешение неоднозначностей

Пример сиамской нейросети - определяем схожесть изображений



похожие пары

НЕпохожие пары

Семантическая близость и разрешение неоднозначностей

Литература

git clone 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

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/