

WORD EMBEDDINGS

КВАНТИТАТИВНЫЙ АНАЛИЗ ТЕКСТА

Кирилл Александрович Маслинский

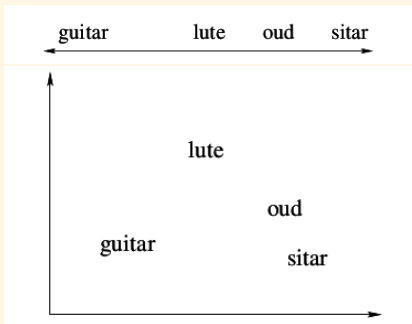
23.05.2022 / 10

НИУ ВШЭ Санкт-Петербург

ПРОСТРАНСТВЕННОЕ МОДЕЛИРОВАНИЕ СЕМАНТИКИ

Schütze

Vector similarity is the only information present in Word Space: semantically related words are close, unrelated words are distant.



Lakoff & Johnson, *Metaphors we live by*, 1980

Метафоры — основа всех абстрактных понятий. Язык вообще и языковое значение в частности построено на метафорах.

Первичные метафоры — непосредственно связаны с телесным опытом.

- Пространство:
 - расположение
 - направление
 - близость (расстояние)

В основе первичные метафоры:

- похожее = близкое
- сущность = место (понятие = место)

Геометрическая метафора языкового значения:

Значения — это точки в семантическом пространстве,
семантическое сходство — расстояние между точками в этом пространстве.

Дистрибутивная гипотеза

Слова со сходными дистрибутивными свойствами обладают сходным значением.

Способы представить дистрибутивное сходство:

- соседствуют с одними и теми же словами
- употребляются в одних и тех же документах
- ...

- LSA/LSI (Latent Semantic Analysis/Indexing) [1988]
- word2vec [2013]
- ELMo/BERT/GPT-2/GPT-3 [2017—2020]

Все они по-разному реализуют одну и ту же идею:

- геометрическая метафора значения
- дистрибутивный метод: построение пространства на основе информации о контекстах слов
- моделируют семантическую близость: смысл в модели имеет только расстояние, но не измерения пространства

В «семантическом пространстве» (word space):

- **Имеет смысл** анализ расстояния между **близкими** по значению словами.
- **Не имеет смысла** анализ расстояния между не связанными по значению словами (удаленными областями в word space).
Что общего между вороном и конторкой?

WORD EMBEDDINGS

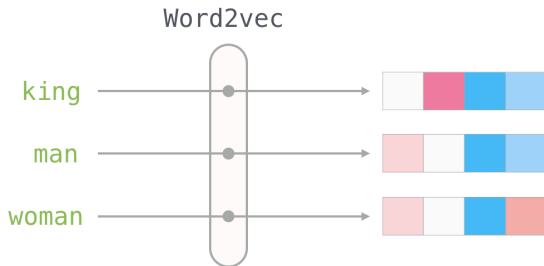
Word embeddings — dense representations of words in a low-dimensional vector space

neural word embeddings word embeddings learned by a neural network

Alternate terms: distributional semantic model/semantic vector space/word space

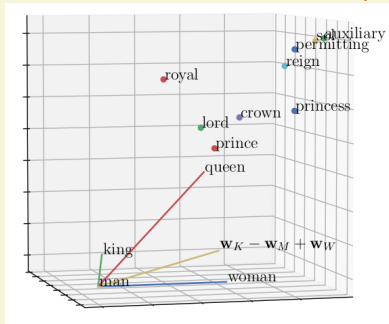
SERENDIPITY:

king - man + woman \approx queen

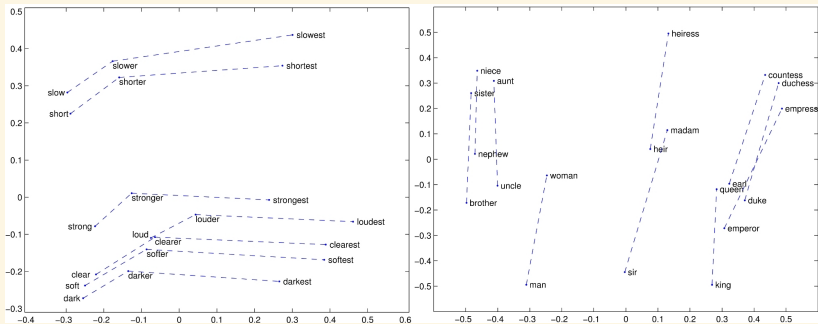


SERENDIPITY:

king – man + woman \approx queen



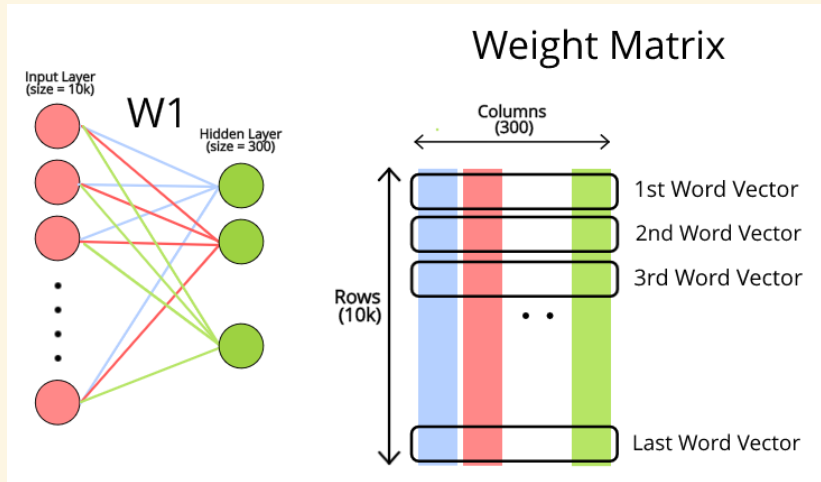
СЕМАНТИЧЕСКАЯ АЛГЕБРА



word2vec Mikolov et al. in 2013

GloVe Pennington et al. 2014

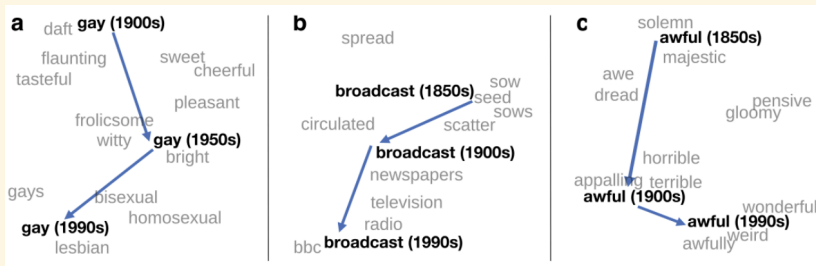
НЕЙРОННЫЕ СЕТИ И DEEP LEARNING REVIVAL



Unsupervised method:

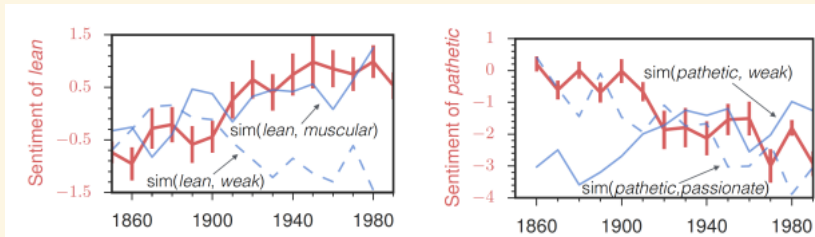
- > большой неаннотированный корпус текстов
- > pre-trained embeddings
- > применение к небольшим массивам размеченных данных

ПРИМЕНЕНИЯ WORD2VEC



Hamilton W. L., Leskovec J., Jurafsky D. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2016. – C. 1489-1501.

ПРИМЕНЕНИЯ WORD2VEC



Hamilton W. L. et al. Inducing domain-specific sentiment lexicons from unlabeled corpora // Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing. – NIH Public Access, 2016. – T. 2016. – C. 595.

SGNS — SKIP-GRAM WITH NEGATIVE SAMPLING AKA WORD2VEC

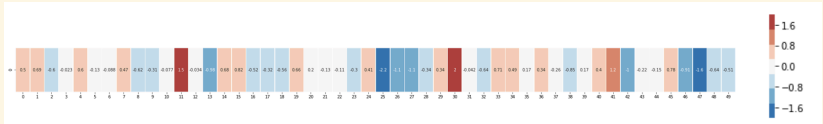
ILLUSTRATIONS FROM:

[HTTP://JALAMMAR.GITHUB.IO/ILLUSTRATED-
WORD2VEC/](http://jalammr.github.io/illustrated-word2vec/)

РУССКИЙ ПЕРЕВОД:

[HTTPS://HABR.COM/RU/POST/446530/](https://habr.com/ru/post/446530/)

EMBEDDINGS: VECTORS FOR WORDS



EMBEDDINGS: VECTORS FOR WORDS

“king”



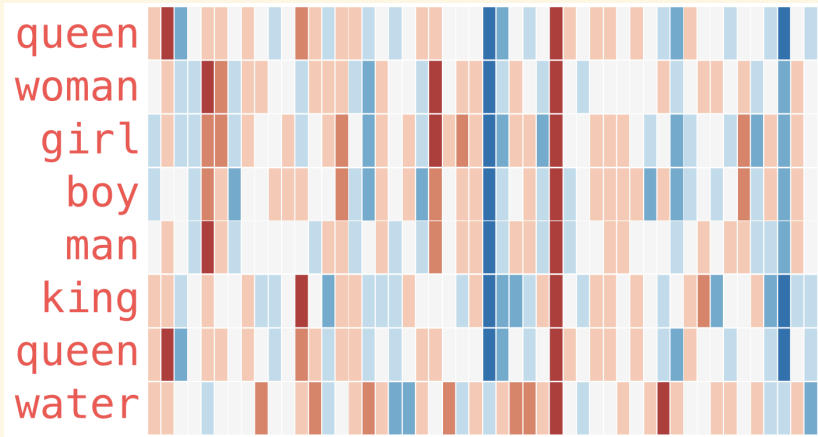
“Man”



“Woman”



EMBEDDINGS: VECTORS FOR WORDS



ЯЗЫКОВАЯ МОДЕЛЬ: ЗАДАЧА ПРЕДСКАЗАНИЯ СЛОВА

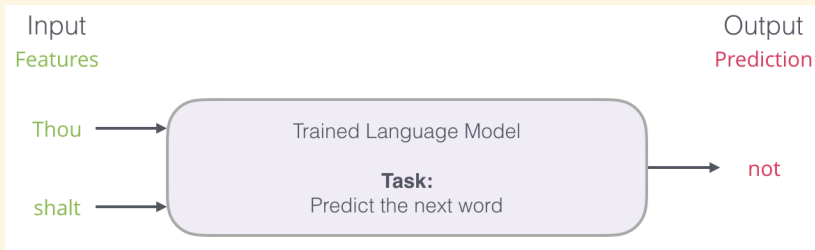
input/feature #1

input/feature #2

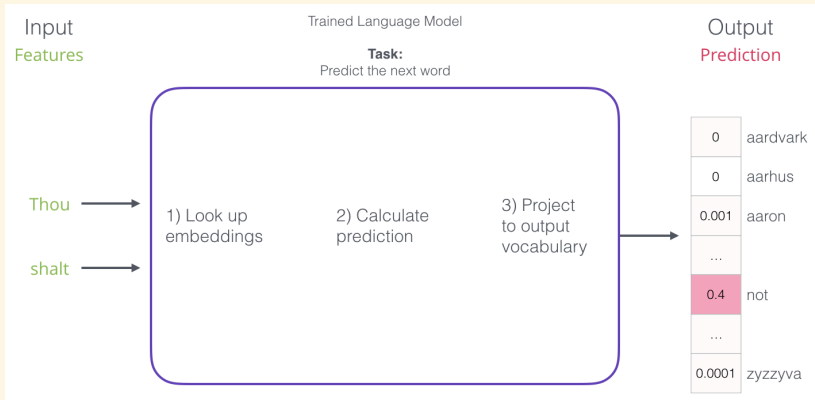
output/label

Thou shalt

ЯЗЫКОВАЯ МОДЕЛЬ: ЗАДАЧА ПРЕДСКАЗАНИЯ СЛОВА



ЯЗЫКОВАЯ МОДЕЛЬ: ЗАДАЧА ПРЕДСКАЗАНИЯ СЛОВА



МЕТОД СКОЛЬЗЯЩЕГО ОКНА

Thou shalt not make a machine in the likeness of a human mind

Sliding window across running text

thou	shalt	not	make	a	machine	in	the	...
------	-------	-----	------	---	---------	----	-----	-----

Dataset

input 1	input 2	output
thou	shalt	not

МЕТОД СКОЛЬЗЯЩЕГО ОКНА

Thou shalt not make a machine in the likeness of a human mind

Sliding window across running text

thou	shalt	not	make	a	machine	in	the	...
thou	shalt	not	make	a	machine	in	the	

Dataset

input 1	input 2	output
thou	shalt	not
shalt	not	make

Jay was hit by a _____ bus in...

by	a	red	bus	in
----	---	-----	-----	----

Jay was hit by a red bus in...

by	a	red	bus	in
----	---	-----	-----	----

input	output
red	by
red	a
red	bus
red	in

Thou shalt not make a machine in the likeness of a human mind

thou	shalt	not	make	a	machine	in	the	...
------	-------	-----	------	---	---------	----	-----	-----

thou	shalt	not	make	a	machine	in	the	...
------	-------	-----	------	---	---------	----	-----	-----

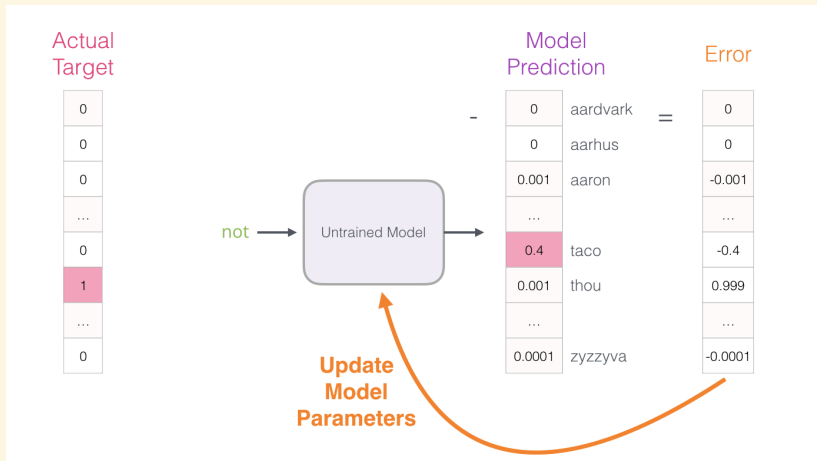
thou	shalt	not	make	a	machine	in	the	...
------	-------	-----	------	---	---------	----	-----	-----

thou	shalt	not	make	a	machine	in	the	...
------	-------	-----	------	---	---------	----	-----	-----

thou	shalt	not	make	a	machine	in	the	...
------	-------	-----	------	---	---------	----	-----	-----

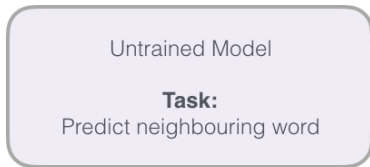
input word	target word
not	thou
not	shalt
not	make
not	a
make	shalt
make	not
make	a
make	machine
a	not
a	make
a	machine
a	in
machine	make
machine	a
machine	in
machine	the
in	a
in	machine
in	the
in	likeness

ОБУЧЕНИЕ МОДЕЛИ



NEGATIVE SAMPLING

not



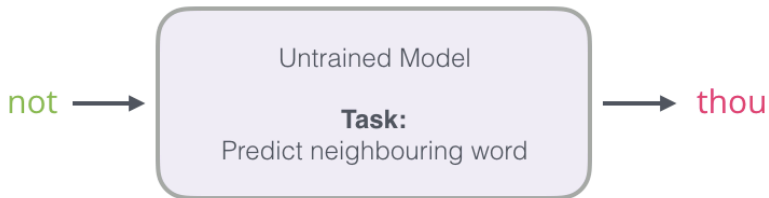
1) Look up
embeddings

2) Calculate
prediction

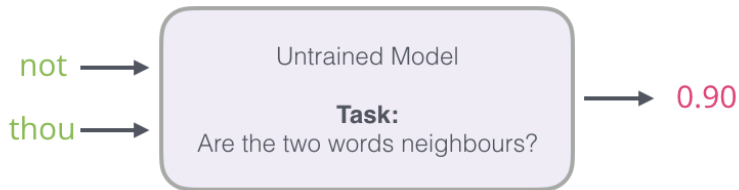
**3) Project
to output
vocabulary**

**[Computationally
Intensive]**

Change Task from



To:



NOISE-CONTRASTIVE ESTIMATION

input word	target word
not	thou
not	shalt
not	make
not	a
make	shalt
make	not
make	a
make	machine

input word	output word	target
not	thou	1
not	shalt	1
not	make	1
not	a	1
make	shalt	1
make	not	1
make	a	1
make	machine	1

not →

thou →

Smartass Model

Task:

Are the two words neighbours?

```
def model(in, out):  
    return 1.0
```

NOISE-CONTRASTIVE ESTIMATION

input word	output word	target
not	thou	1
not		0
not		0
not	shalt	1
not	make	1

 Negative examples

NOISE-CONTRASTIVE ESTIMATION

Pick randomly from vocabulary
(random sampling)

input word	output word	target
not	thou	1
not	aaron	0
not	taco	0
not	shalt	1
not	make	1

Word	Count	Probability
aardvark		
aarhus		
aaron		
taco		
thou		
zyzzyva		

SGNS: SKIPGRAM WITH NEGATIVE SAMPLING

Skipgram

shalt	not	make	a	machine
-------	-----	------	---	---------

input	output
make	shalt
make	not
make	a
make	machine

Negative Sampling

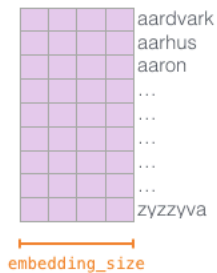
input word	output word	target
make	shalt	1
make	aaron	0
make	taco	0

ОБУЧЕНИЕ МОДЕЛИ WORD2VEC (SGNS)

Embedding



Context



Embedding

			aardvark
			aarhus
			aaron
			...
			not
			...
			...
			...
			zyzzyva

Context

			aardvark
			aarhus
			aaron
			...
			taco
			...
			thou
			...
			zyzzyva

Look up
embeddings



not



aaron











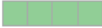



taco









thou









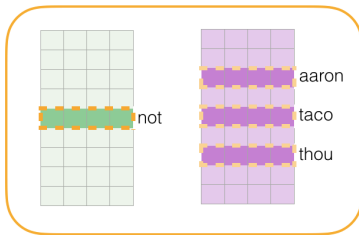
input word	output word	target	input • output
not 	thou 	1	0.2
not 	aaron 	0	-1.11
not 	taco 	0	0.74

input word	output word	target	input • output	sigmoid()
not 	thou 	1	0.2	0.55
not 	aaron 	0	-1.11	0.25
not 	taco 	0	0.74	0.68

input word	output word	target	input • output	sigmoid()	Error
not 	thou 	1	0.2	0.55	0.45
not 	aaron 	0	-1.11	0.25	-0.25
not 	taco 	0	0.74	0.68	-0.68

ОБУЧЕНИЕ WORD2VEC

input word	output word	target	input • output	sigmoid()	Error
not 	thou 	1	0.2	0.55	0.45
not 	aaron 	0	-1.11	0.25	-0.25
not 	taco 	0	0.74	0.68	-0.68



Update
Model
Parameters

ПАРАМЕТРЫ МОДЕЛИ WORD2VEC

Window size: 5



Window size: 15



КОЛИЧЕСТВО ОТРИЦАТЕЛЬНЫХ ПРИМЕРОВ

Negative samples: 2

input word	output word	target
make	shalt	1
make	aaron	0
make	taco	0

Negative samples: 5

input word	output word	target
make	shalt	1
make	aaron	0
make	taco	0
make	finglonger	0
make	plumbus	0
make	mango	0

WORD EMBEDDINGS AND DISTRIBUTIONAL SEMANTIC MODELS

СРАВНЕНИЕ С ДИСТРИБУТИВНЫМИ МОДЕЛЯМИ

модели	контекст	тип отношения	пример
LSA, pLSA, LDA	документ	semantic relatedness	boat — water
word embeddings, HAL, Random indexing, BEAGLE	слова	semantic similarity	boat — ship

КАК МЫ И ДУМАЛИ, НИКАКОЙ РАЗНИЦЫ!

levy2014neural

WORD2VEC: СЕКРЕТ УСПЕХА

- Dynamic context window: $decay = 1/distance$
- Subsampling frequent words: randomly delete words that are too common
- Deleting rare words

- Shifted PMI $SPPMI(w, c) = \max(pmi(w, c) - \log(k), 0)$
- Context distribution smoothing

$$pmi(w, c) = \log \frac{p(w, c)}{p(w)p_{\alpha}(c)}, \text{ where } p_{\alpha}(c) = \frac{f(c)^{\alpha}}{\sum_c f(c)^{\alpha}}, \alpha = \frac{3}{4}$$

- Hyperparameters vs. algorithms:
Hyperparameter settings are often more important than algorithm choice. No single algorithm consistently outperforms the other methods.
- Hyperparameters vs. more data:
Training on a larger corpus helps for some tasks. In 3 out of 6 cases, tuning hyperparameters is more beneficial.

DEBUNKING PRIOR CLAIMS

1. Are embeddings superior to distributional methods?
With the right hyperparameters, no approach has a consistent advantage over another.
2. Is GloVe superior to SGNS?
SGNS outperforms GloVe on all tasks.
3. Is CBOW a good word2vec configuration?
CBOW does not outperform SGNS on any task.

EXPLAINING WORD EMBEDDINGS

POINTWISE MUTUAL INFORMATION

PMI

$$pmi(x; y) = \log \frac{p(x, y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)}$$

Positive PMI

$$ppmi(x; y) = \max(pmi(x; y), 0)$$

ЧЕТЫРЕ ВИДА СХОДСТВА

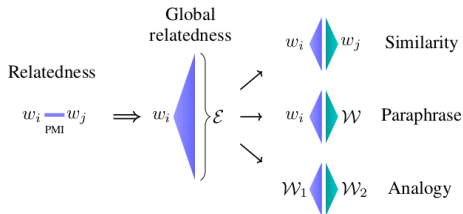


Figure 2: Interconnection between semantic relationships: relatedness is a base pairwise comparison (measured by PMI); *global relatedness* considers relatedness to all words (PMI vector); similarity, paraphrase and analogy depend on global relatedness between words ($w \in \mathcal{E}$) and word sets ($\mathcal{W} \subseteq \mathcal{E}$).

¹Allen, C., Balazevic, I., & Hospedales, T. (2019). What the vec? towards probabilistically grounded embeddings. Advances in Neural Information Processing Systems, 32, 7467-7477.