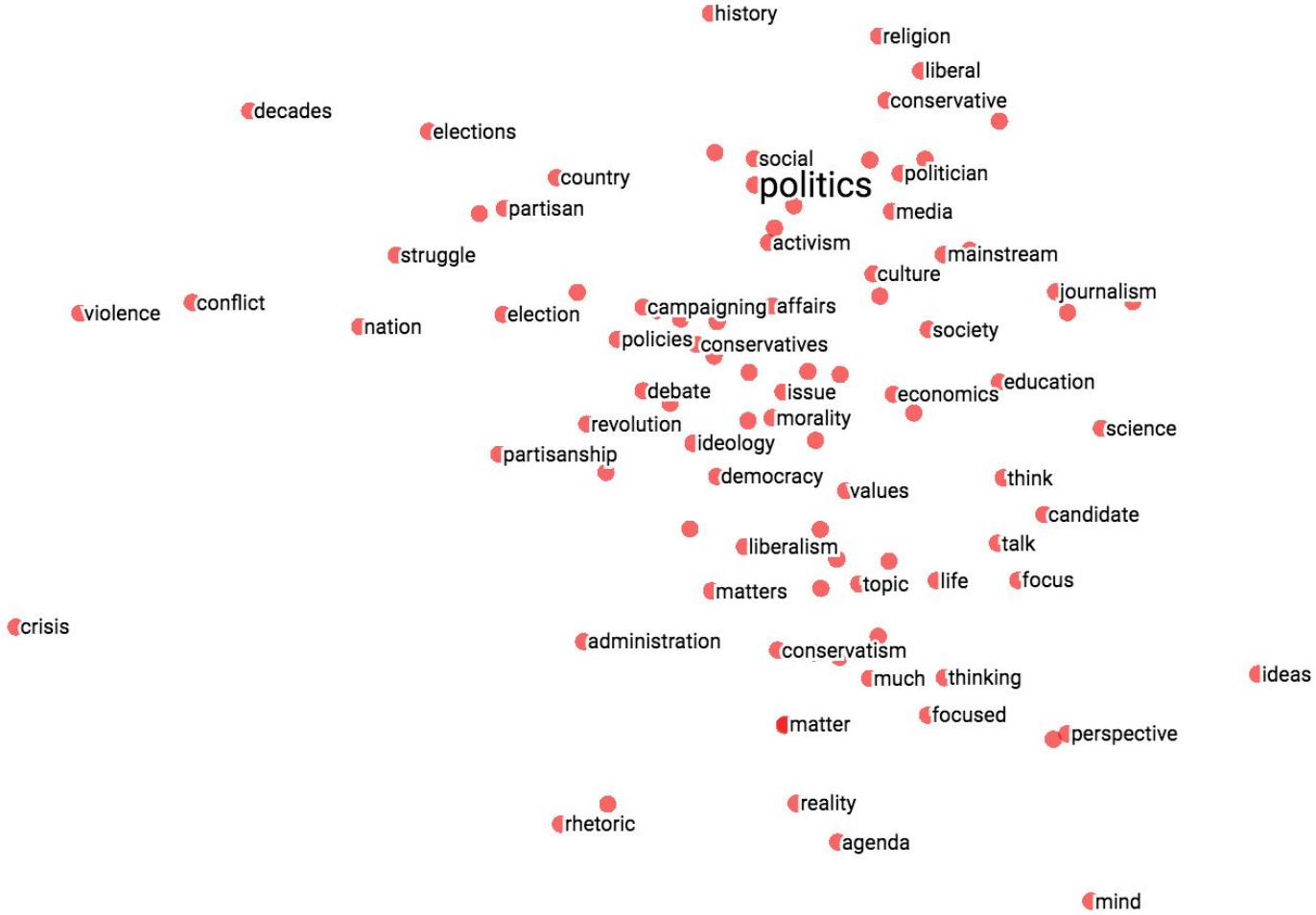
Embeddings

DMIA


history
religion
liberal
conservative
decades
elections
social
politician
country
politics
partisan
media
activism
struggle
mainstream
culture
journalism
violence
conflict
election
campaigning
affairs
nation
society
policies
conservatives
debate
issue
economics
education
revolution
morality
science
ideology
partisanship
democracy
think
values
candidate
liberalism
talk
topic
life
focus
matters
crisis
administration
conservatism
much
thinking
ideas
matter
focused
perspective
reality
rhetoric
agenda
mind

Зачем это нужно?

1) Классификация текстов
2) Кластеризация текстов
3) Машинный перевод

Нужно как-то представить вектора слов
Ваши идеи?

Нужно как-то представить вектора слов
Ваши идеи?

One hot encoding

Хочу домашку по трендам

	Хочу	домашку	по	трендам
0	1	0	0	0
1	0	1	0	0
2	0	0	1	0
3	0	0	0	1

Нужно как-то представить вектора слов
Ваши идеи?

One hot encoding

Хочу домашку по трендам

	Хочу	домашку	по	трендам
0	1	0	0	0
1	0	1	0	0
2	0	0	1	0
3	0	0	0	1

1) Не учитывается близость между словами
2) Опечатки-ошибки игнорируются
3) Слишком много признаков

Bag of words


BIG
DATA
INFORMATION
STORAGE
LARGE
SOFTWARE
SAN
NAS
SETS
MILLION
SCIENCE
VOLUME
GOVERNMENT
TECHNOLOGIES
PROCESS
SYSTEMS
MANAGEMENT
DATABASES
PROCESSING
SIZE
EXABYTES
PETABYTES
SEARCH
ANALYSIS
RESEARCH
LEARNING
WORLD
EVERY
BASED
INTERNET
TIMES
DEVELOPMENT
TECHNOLOGY
NEW
BILLION
PARALLEL
FUTURE
CRITIQUES
INITIATIVE
ECONOMIC
TERABYTES
NATIONAL
INTELLIGENCE
APPROACHES
AMOUNT
REAL
APPLICATIONS
INFRASTRUCTURE
ANALYTICS
FRAMEWORK
BETWEEN
BUSINESS
SOCIAL
YEARS
RESULTS
NETWORKS
HYPOTHESIS
PAST
COMPLEX
CHALLENGES
PRIVACY
INTERNATIONAL
IMPLEMENTATION
BECOMES
PARADIGM
FUNDING
INSIGHT
TOOLS
RECORDS
COST
DISTRIBUTED
RELEVANT
USED
ARCHITECTURE
SUCCESSFUL
TERMS
YEAR
SHARED

Что такое близкие слова?

Скажи мне, кто твой сосед и я скажу тебе, кто ты

Будем смотреть не только на слова, но и на их контексты!

Нужна большая общность!

Матрица слово-контекст

	и	машинное	обучение	рудн	иностранец
и	1000	40	50	2	3
машинное		100	45	0	0
обучение			150	3	5
рудн				5	10
иностранец					30

Нужна большая общность!

Матрица слово-контекст

	и	машинное	обучение	рудн	иностранец
и	1000	40	50	2	3
машинное		100	45	0	0
обучение			150	3	5
рудн				5	10
иностранец					30

Использовать не количество а что-то похитрей

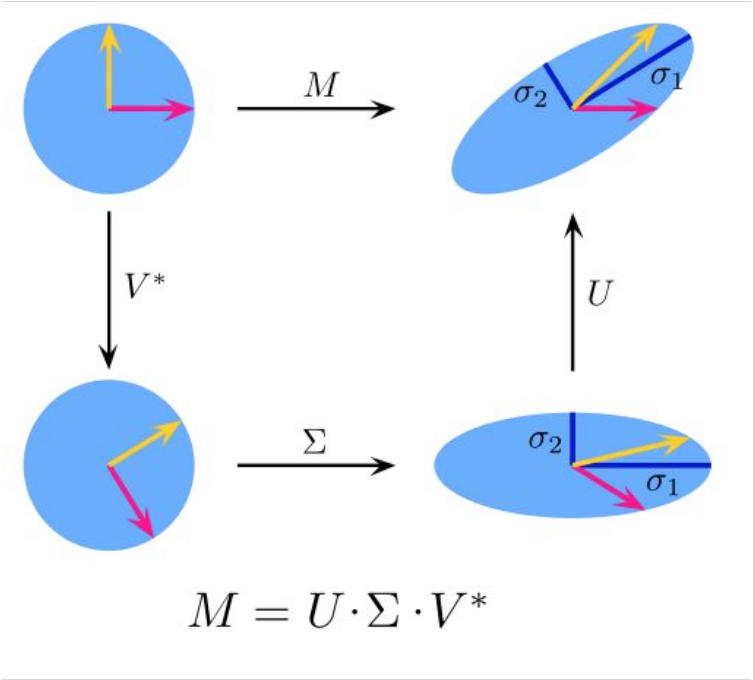
$$\mathrm{PMI}(w_1, w_2) = \log_2 \frac{P(w_1 w_2)}{P(w_1)P(w_2)}$$

SVD- разложение

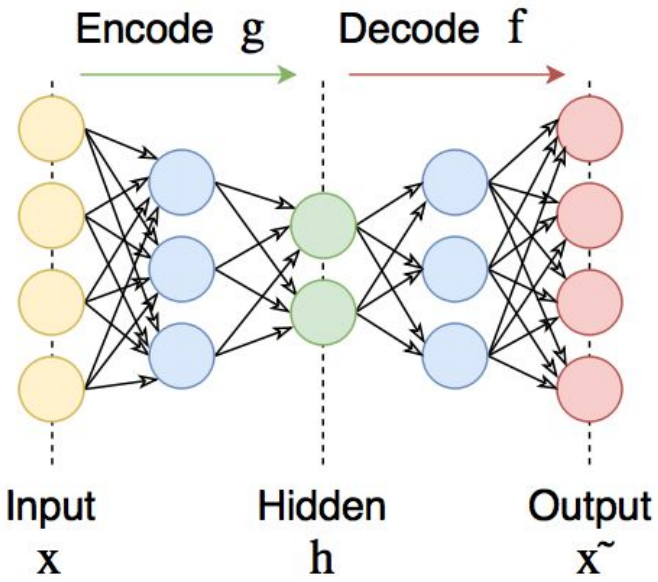
$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$$

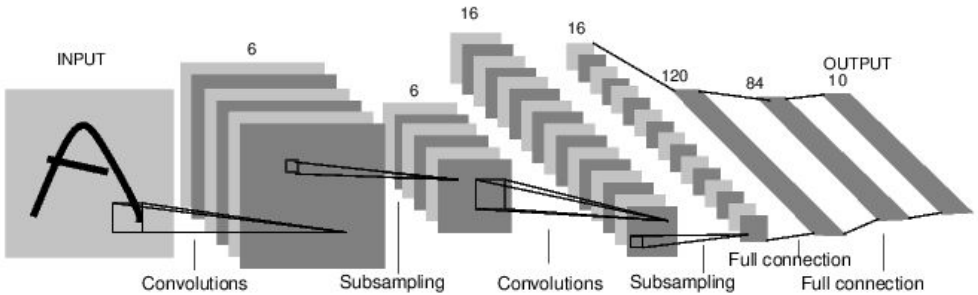
U- матрица слов

V - матрица контекстов



Следующий уровень - нейронные сети

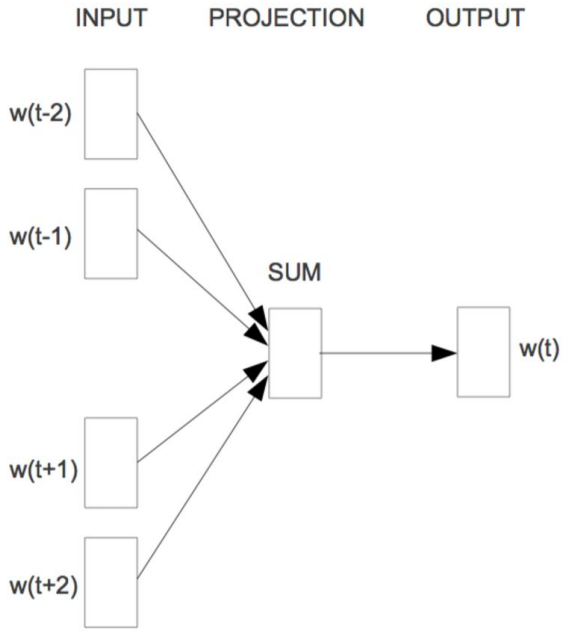




Word2vec

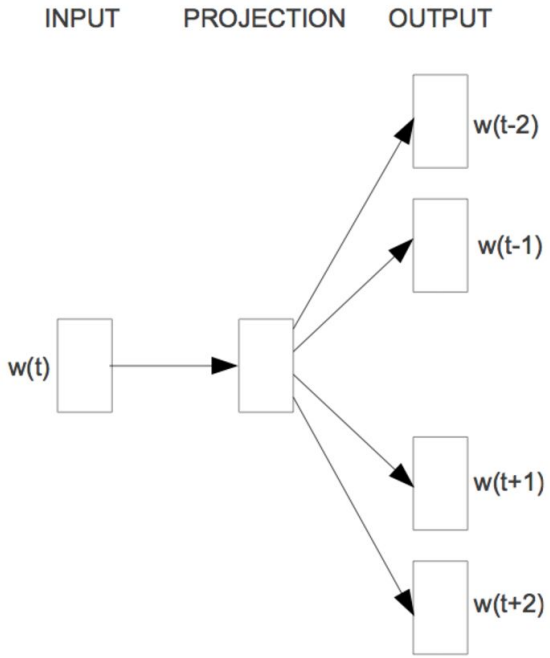
Идея - давайте учить сеть предсказывать вероятности встречаемость слова в контексте и надеятся, что признаки, полученные при обучении окажутся хорошими представлениями для слов

CBOW



$$J_\theta = \frac{1}{T} \sum_{t=1}^{T} \log p(w_t \mid w_{t-n}, \cdots, w_{t-1}, w_{t+1}, \cdots, w_{t+n}).$$

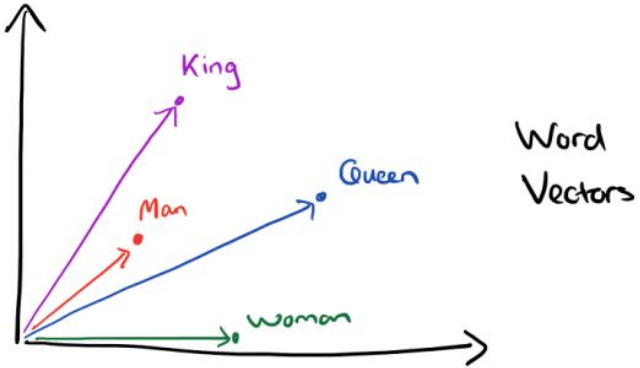
Skip-gram



$$J_\theta = \frac{1}{T} \sum_{t=1}^{T} \sum_{-n \le j \le n, \neq 0} \log p(w_{t+j} \mid w_t).$$

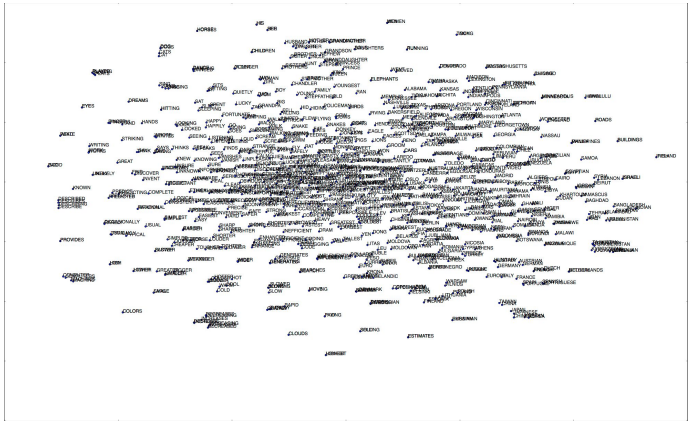
Свойства

Линейность



Неинтерпретируемость

компонент



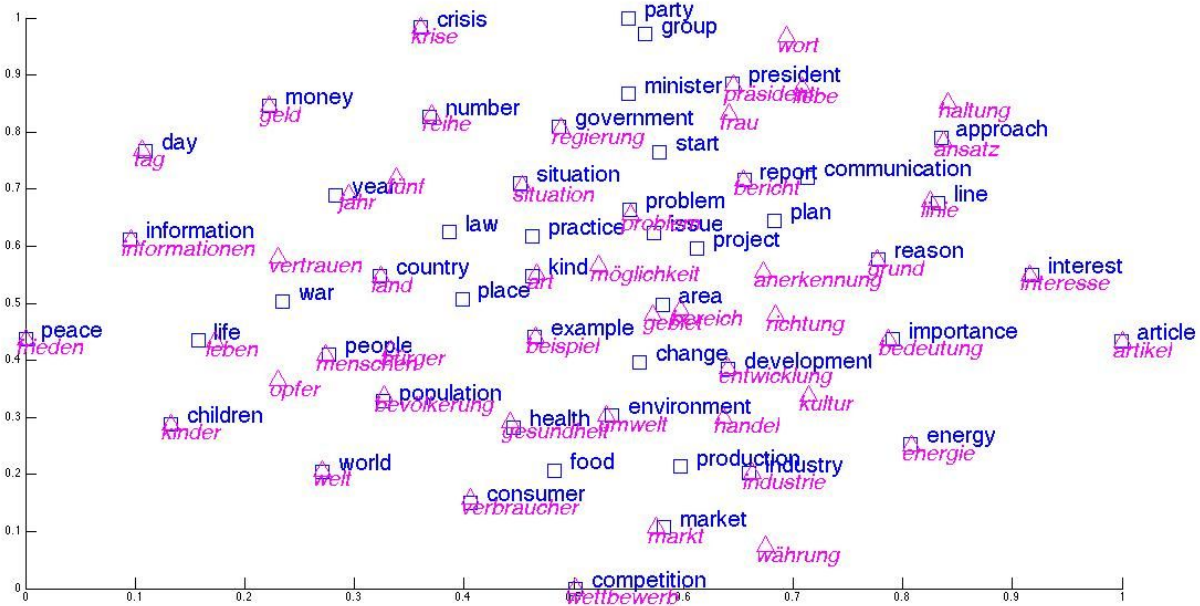
FastText

Идея - давайте расматривать не слова, а n- граммы!

В чем плюсы FastText:

1) Экономно
2) Хорошо работает на редких словах (идеально для опечаток)
3) Работает на словах, которых не было в обучении

Кросс-эмбединги



В чем смысл?

1) Отобразить слова из разных языков в одно пространство
2) Эффективно решать задачу машинного перевода
3) Использовать информацию из одного языка в другом

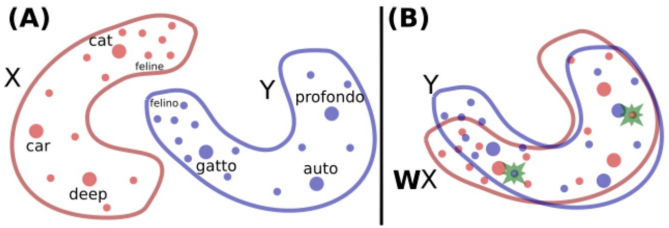
Когда есть учитель

1) Берем 2 набора embeddings, обученных на двух разных языках независимо, и перевод слов с одного языка на другой
2) Найдем отображение одного пространства эмбеддингов в другое так, чтобы расстояние между парами было наименьшим

Процесс отображения

$$W^\star = \underset{W \in M_d(\mathbb{R})}{\operatorname{argmin}} \|WX - Y\|_{\mathrm{F}} \qquad (1)$$

d - dimension of the embeddings
$M_d(R)$ - space of d × d matrices of real numbers
X and Y - two aligned matrices of size d × n containing the embeddings of the words in the parallel vocabulary.
W - линейное отображение между языковыми пространствами



Обучение без учителя

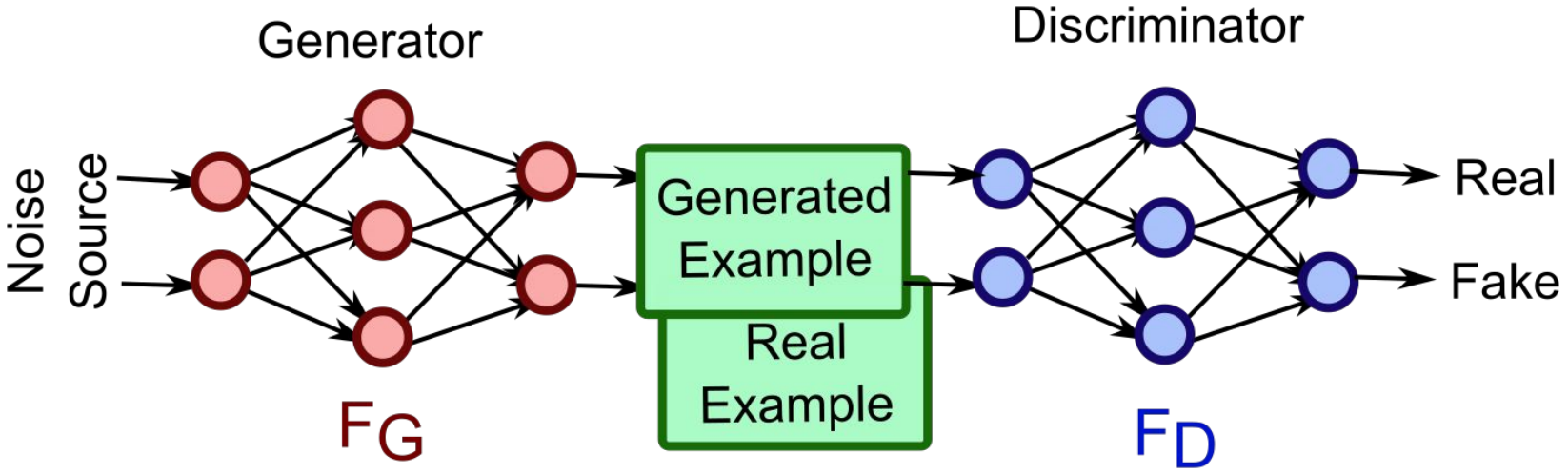
Делаем то же самое без обучающей выборки

1) Применяем advertising learning
2) Уточняем матрицу перехода по самым лучшим переводам
3) Ищем ближайших соседей
4) Превосходим по качеству supervised алгоритмы

Флэшбек

Флэшбек

GAN - та же идея!



$\mathcal{X} = \{x_1, ..., x_n\}$ and $\mathcal{Y} = \{y_1, ..., y_m\}$ - два набора эмбеддингов

θ_D - параметры дискриминатора

W - отображение

Discriminator loss

$$\mathcal{L}_D(\theta_D|W) = -\frac{1}{n}\sum_{i=1}^{n}\log P_{\theta_D}\left(\text{source} = 1\middle|Wx_i\right) - \frac{1}{m}\sum_{i=1}^{m}\log P_{\theta_D}\left(\text{source} = 0\middle|y_i\right). \quad (3)$$

Mapper loss

$$\mathcal{L}_W(W|\theta_D) = -\frac{1}{n}\sum_{i=1}^{n}\log P_{\theta_D}\left(\text{source} = 0\middle|Wx_i\right) - \frac{1}{m}\sum_{i=1}^{m}\log P_{\theta_D}\left(\text{source} = 1\middle|y_i\right). \quad (4)$$

Результаты

	English to italian			Italian to english		
	P@1	P@5	P@10	P@1	P@5	P@10
Methods with cross-lingual supervision						
Mikolov et al. (2013b) †	10.5	18.7	22.8	12.0	22.1	26.7
Dinu et al. (2015) †	45.3	72.4	80.7	48.9	71.3	78.3
Smith et al. (2017) †	54.6	72.7	78.2	42.9	62.2	69.2
Procrustes - NN	42.6	54.7	59.0	53.5	65.5	69.5
Procrustes - CSLS	**66.1**	77.1	80.7	**69.5**	**79.6**	**83.5**
Methods without cross-lingual supervision						
Adv - CSLS	42.5	57.6	63.6	47.0	62.1	67.8
Adv - Refine - CSLS	65.9	**79.7**	**83.1**	69.0	**79.7**	83.1

Table 3: English-Italian sentence translation retrieval. We report the average P@k from 2,000 source queries using 200,000 target sentences. We use the same embeddings as in Smith et al. (2017). Their results are marked with the symbol †.

Вывод

Эмбеддинги - это здорово!