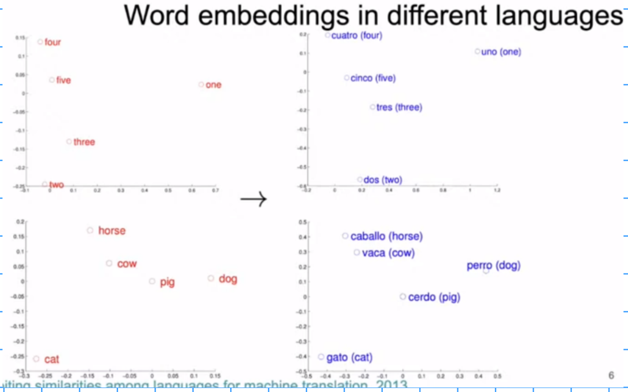
Как можно получить переводчик без больших параллельных корпусов данных?

предпосылка: Языки народов, живущих в похожих условиях, похожи. Потому что язык описывает мир и вз-я об-в между собой. Получается слова с одним смыслом встречаются в похожем контексте.



iting similarities among languages for machine translation, 2013

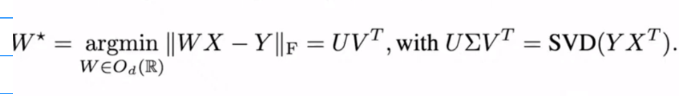
6

идея: давайте просто отобразим облако точек из 1-го пр-ва в другое, выучив какое-нибудь линейное преобразование

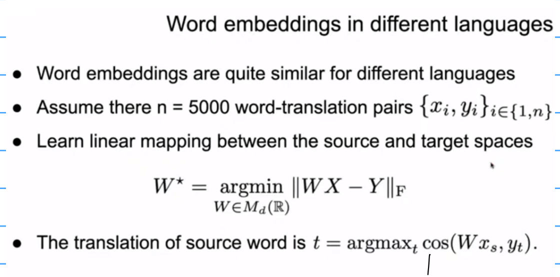
1) берём n пар слов с 2-х языков и обучаем на них матрицу перехода

2) применяем на любое слово и ищем ближайшее

3) а давайте сделаем W ортогональной



$$W^\star = \underset{W \in O_d(\mathbb{R})}{\operatorname{argmin}} \|WX - Y\|_F = UV^T, \text{with } U\Sigma V^T = \operatorname{SVD}(YX^T).$$



Word embeddings in different languages

- Word embeddings are quite similar for different languages
- Assume there n = 5000 word-translation pairs $\{x_i, y_i\}_{i \in \{1,n\}}$
- Learn linear mapping between the source and target spaces

$$W^\star = \underset{W \in M_d(\mathbb{R})}{\operatorname{argmin}} \|WX - Y\|_F$$

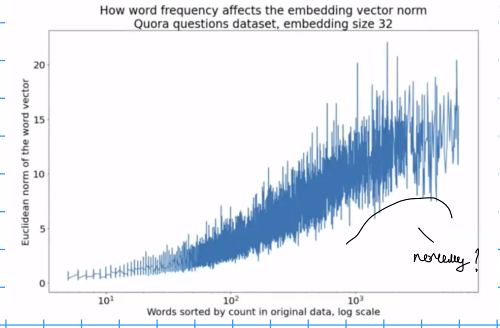
- The translation of source word is $t = \operatorname{argmax}_t \cos(Wx_s, y_t)$.

почему cos, а не что-то другое?

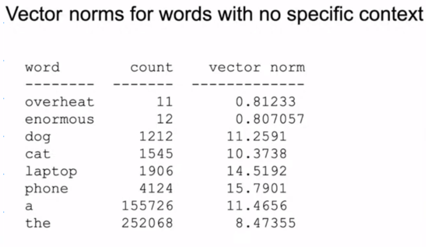
cos – показывает сонаправленность векторов и не смотрит на их норму

пример: (при подходах типа BoW) тексты одной тематики, но разного объёма будут иметь разные нормы эмб-в, но сонаправлены. а норма не сонаправленного эмб-га текста другой тематики, но близкий по норме близки

норма вектора в зав-ти от частоты встр-ти слова



почему?



Vector norms for words with no specific context

```
word        count    vector norm
--------   -------   -------------
overheat        11        0.81233
enormous        12        0.807057
dog           1212       11.2591
cat           1545       10.3738
laptop        1906       14.5192
phone         4124       15.7901
a           155726       11.4656
the         252068        8.47355
```

ответ: общеупотребимые слова встречаются в разных контекстах. Поэтому при обучении антиградиент дёргает их в разные стороны и не уносит далеко от (0, 0, ..., 0)

Поэтому cos лучше использовать и с ембеддингами

Свёртки в работе с текстами

Идея: улавливать локальный контекст, проходясь фильтрами (окнами)

Что делать с разной длиной входа?

1) RNN

2) адаптивный пулинг

— CNN лучше ембеддят фрагменты текста, который редко бывает без префикса