9

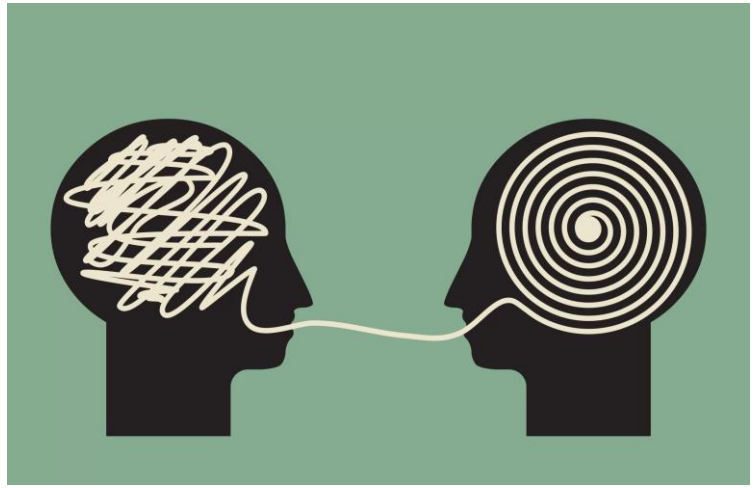
Введение в NLP

word2vec

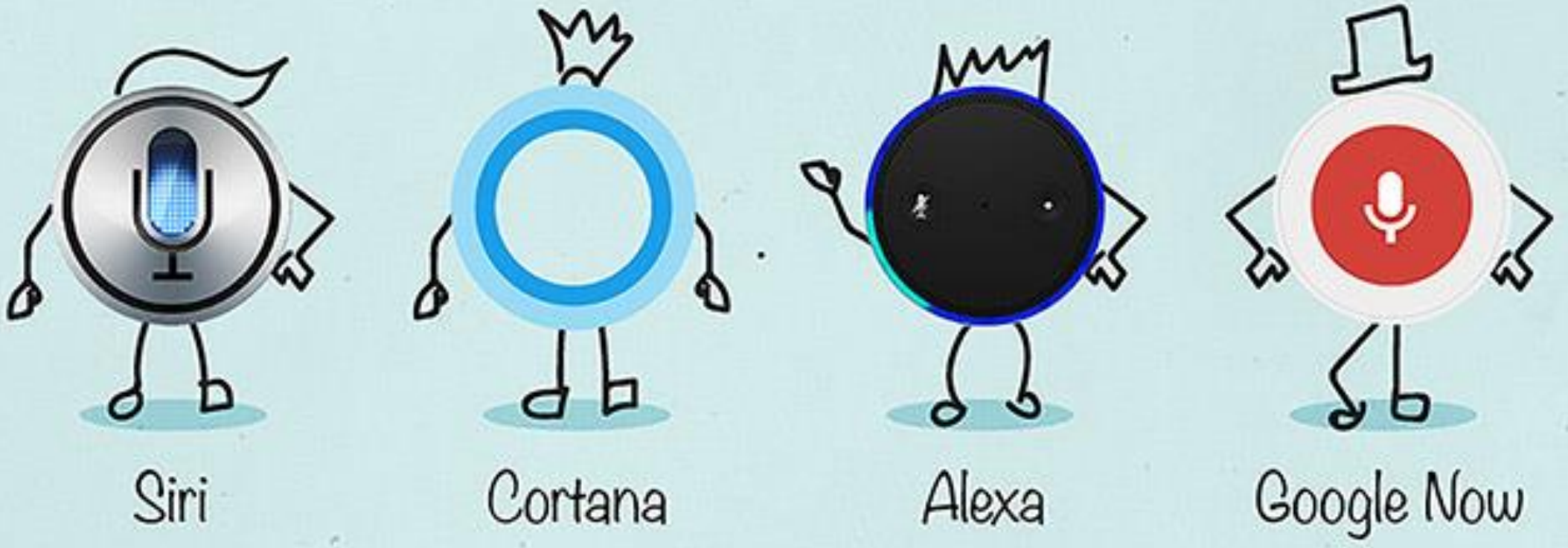
Обработка естественного языка Natural Language Processing (NLP)

Обработка естественного языка (*Natural Language Processing, NLP*) — общее направление искусственного интеллекта и математической лингвистики. Оно изучает проблемы компьютерного анализа и синтеза естественных языков.

Wikipedia



東 = est = ESTE = **east**
西 = ouest = OESTE = **west**
北 = nord = NORTE = **north**
南 = sud = SUR = **south**


Siri
Cortana
Alexa
Google Now

Почему это вообще сложно?

Он видел их семью своими глазами

Эти типы стали есть в цехе

Habr

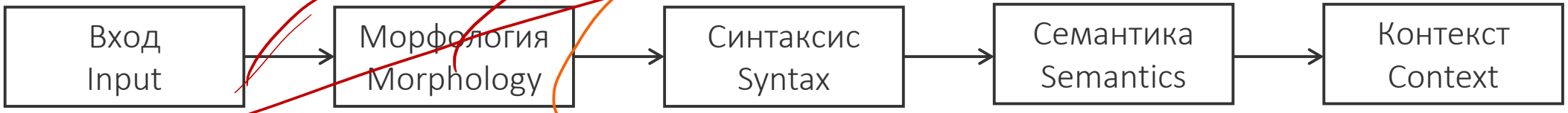
парень 25 лет ищет подработку на субботу и воскресенье

новые куртки есть размеры от 2000 рублей

Перед нами стол.
На столе стакан и вилка.
Что они делают?
Стакан стоит, а вилка лежит.
Если мы воткнем вилку в столешницу, вилка будет стоять.
То есть стоят вертикальные предметы, а лежат горизонтальные?
Добавляем на стол тарелку и сковороду.
Они вроде горизонтальные, но на столе стоят.
Теперь положим тарелку в сковородку.
Там она лежит, а ведь на столе стояла.

Livejournal

NLP pipeline



Deep learning

Например:

grammarly

Текст Распознавание речи OCR	Проверка текста Поиск по словам	Крутая проверка текста Парсинг (+crawling)	Машинный перевод Sentiment analysis	Чат-боты AI Ассистенты

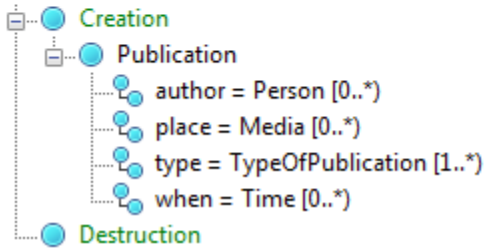
Google Yandex Bing

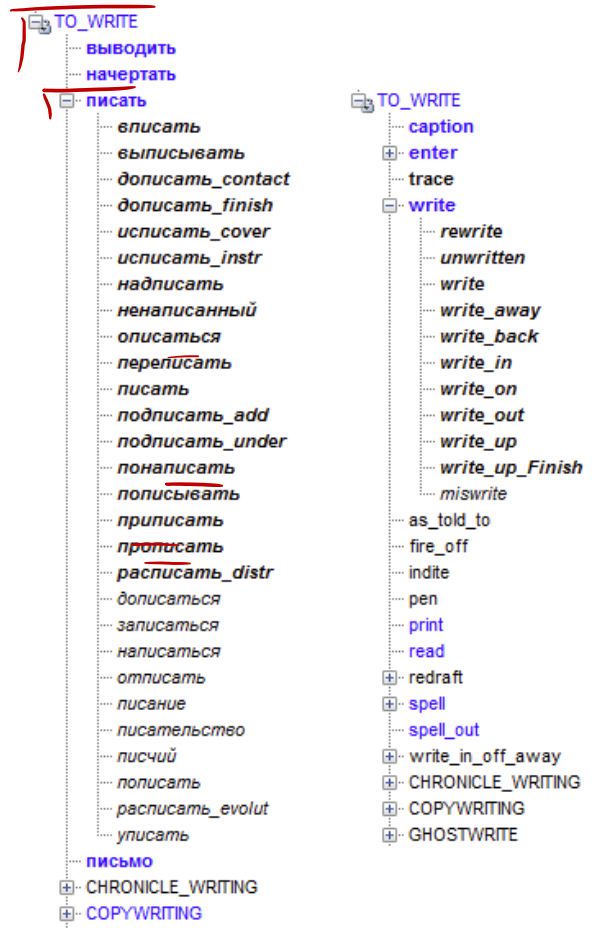
Генерация текста, summarization, итд итд

Как это делают без нейросетей

Онтоинженер Даня в 2014 году написал пост на Хабр

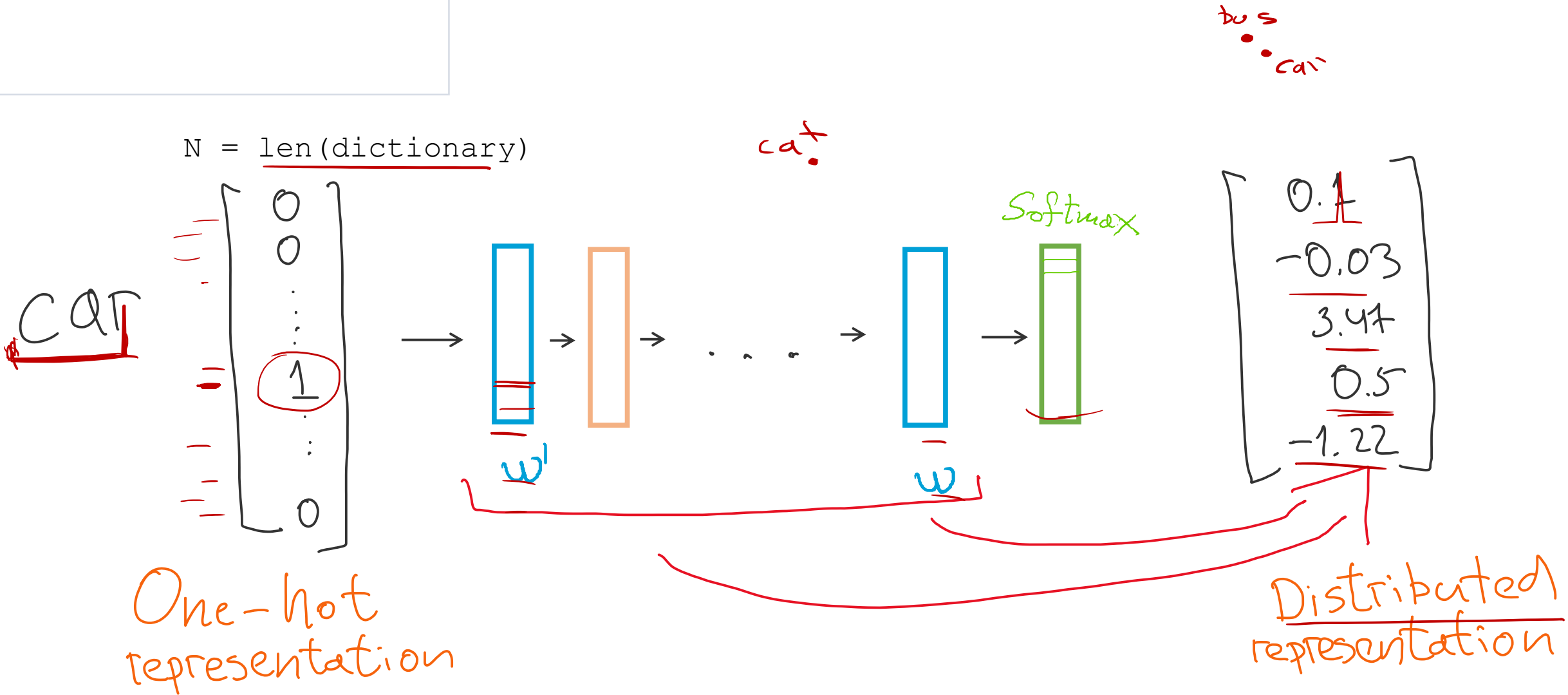
Для этого потребуется модель предметной области (онтология) и правила извлечения информации. Созданием онтологий и правил занимается специальный отдел компьютерных лингвистов, которых мы называем онтоинженерами. Пример онтологии, моделирующей факт публикации:

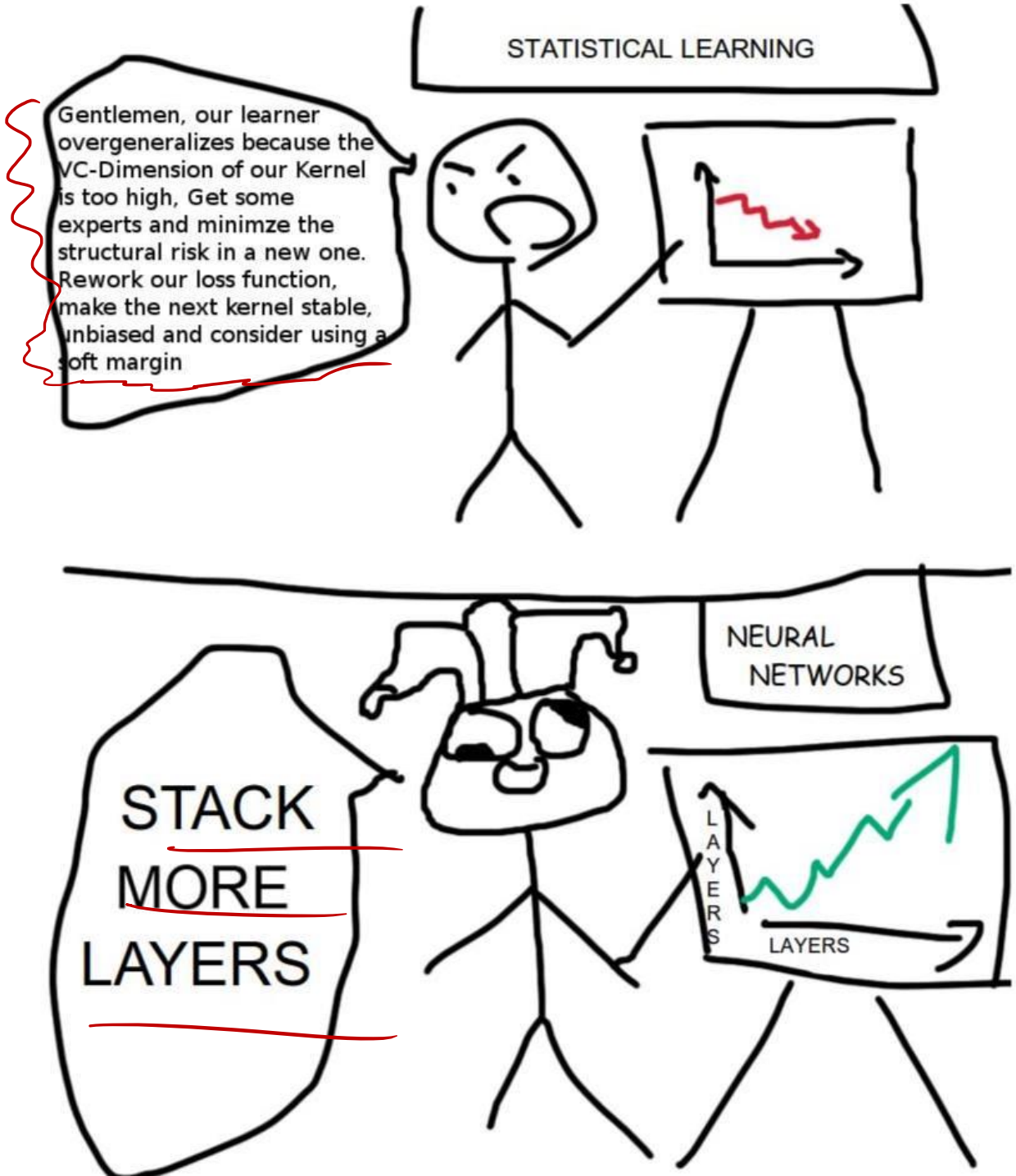




[Алгоритм извлечения информации в ABBYY Compreno]

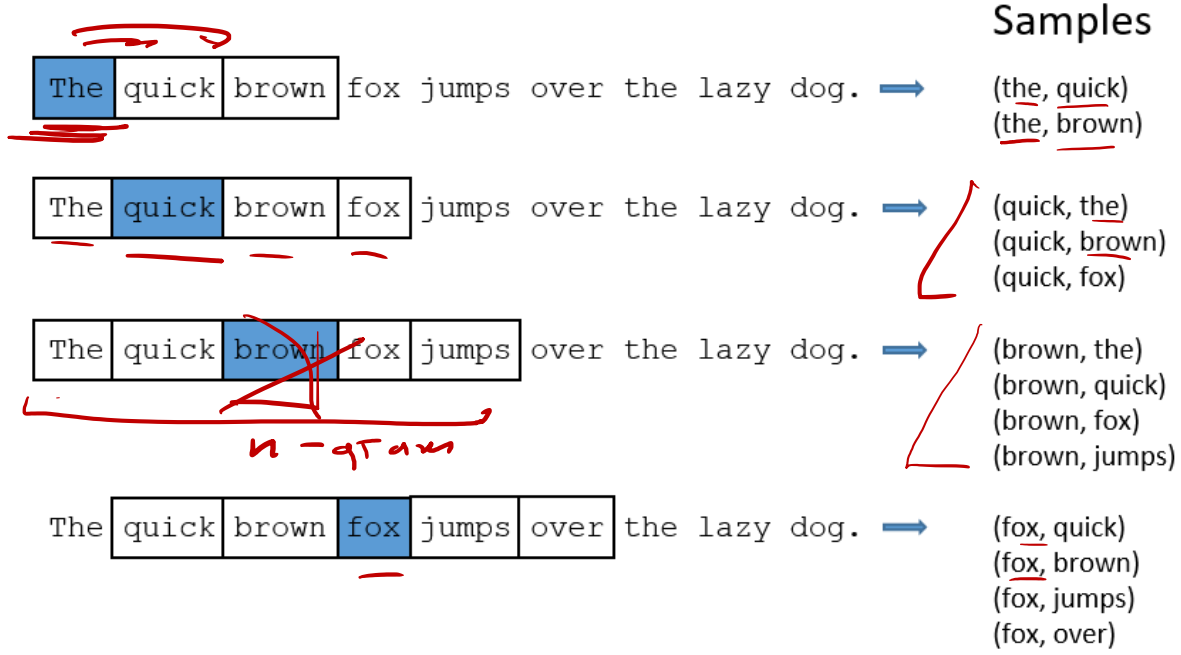
Deep NLP
Из символьного в непрерывное




STATISTICAL LEARNING
Gentlemen, our learner overgeneralizes because the VC-Dimension of our Kernel is too high, Get some experts and minimze the structural risk in a new one. Rework our loss function, make the next kernel stable, unbiased and consider using a soft margin
NEURAL NETWORKS
STACK MORE LAYERS
LAYERS
LAYERS

word2vec

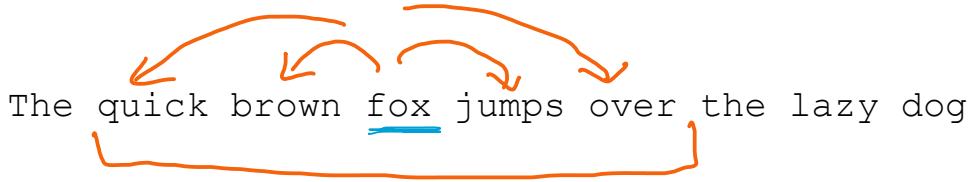
The quick brown fox jumps over the lazy dog



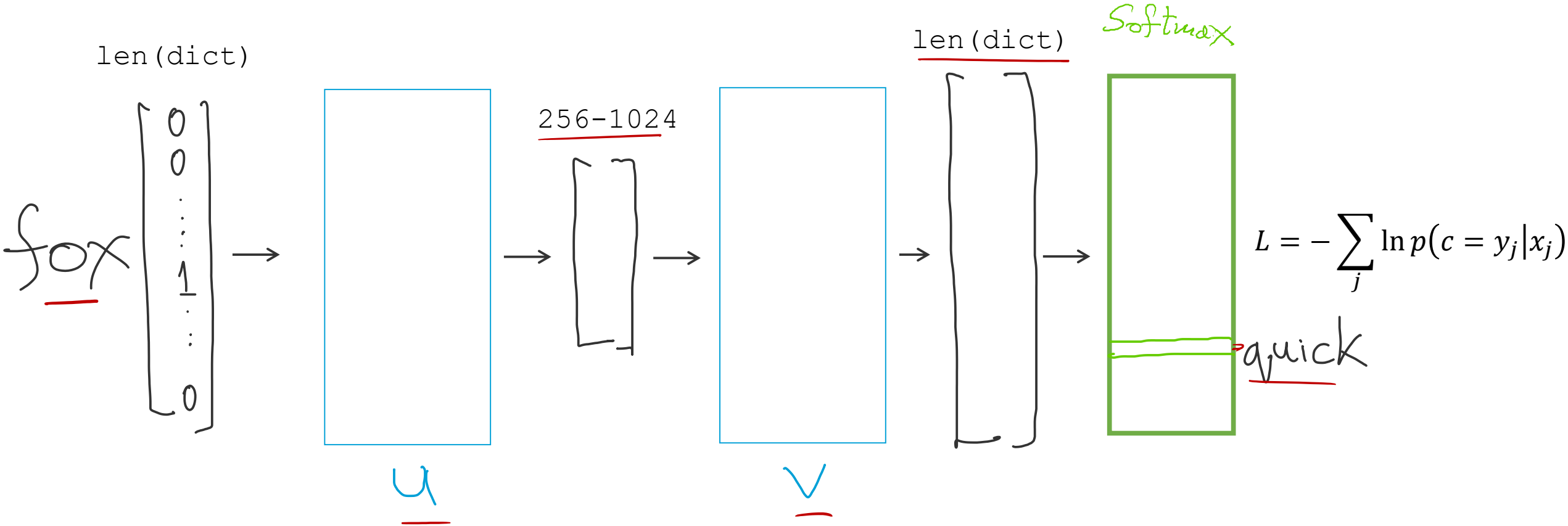
Skip-gram

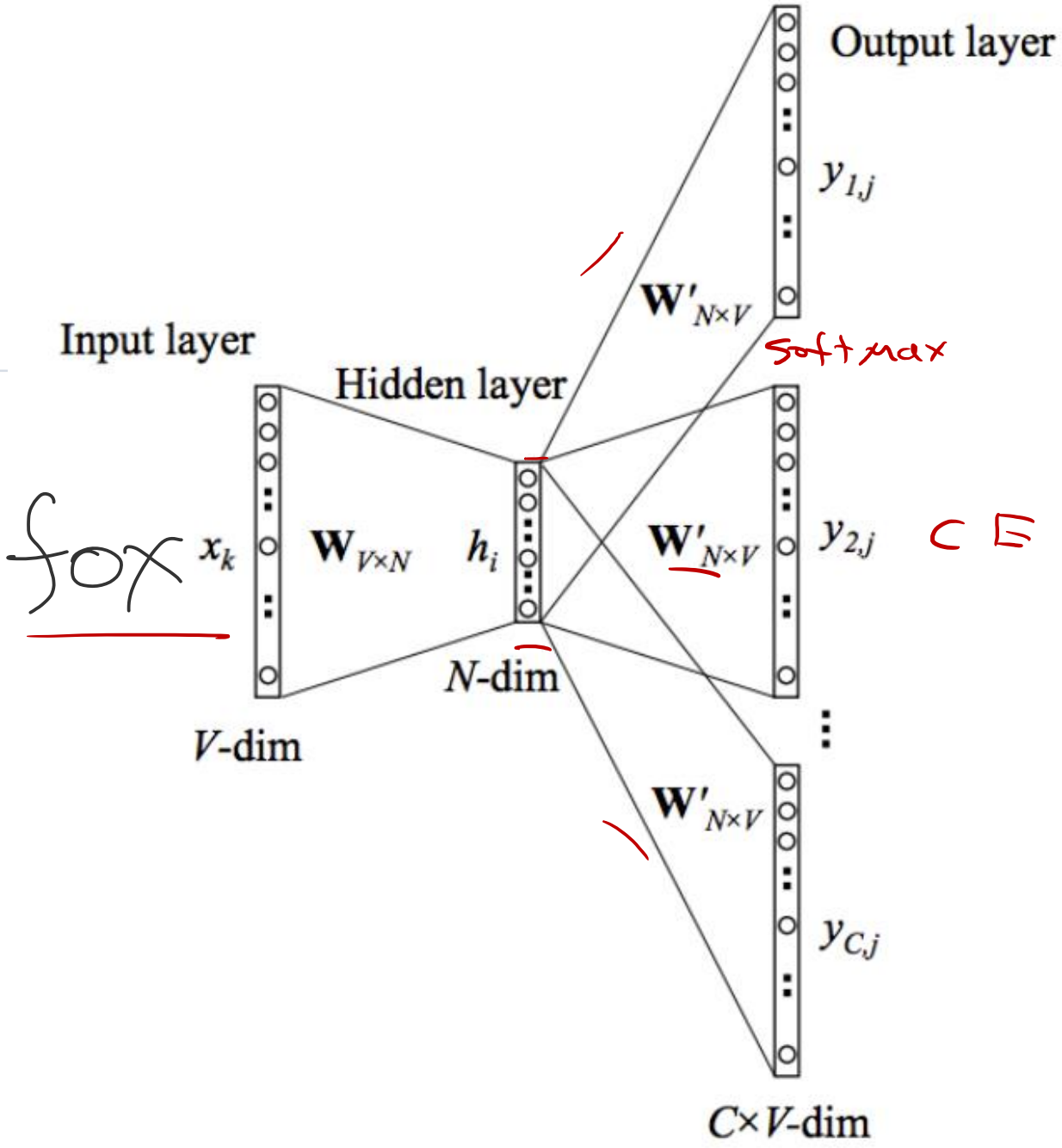
[Word2Vec Tutorial - The Skip-Gram Model]

word2vec




```
fox -> quick
fox -> brown
fox -> jumps
fox -> over
```





quick

Softmax

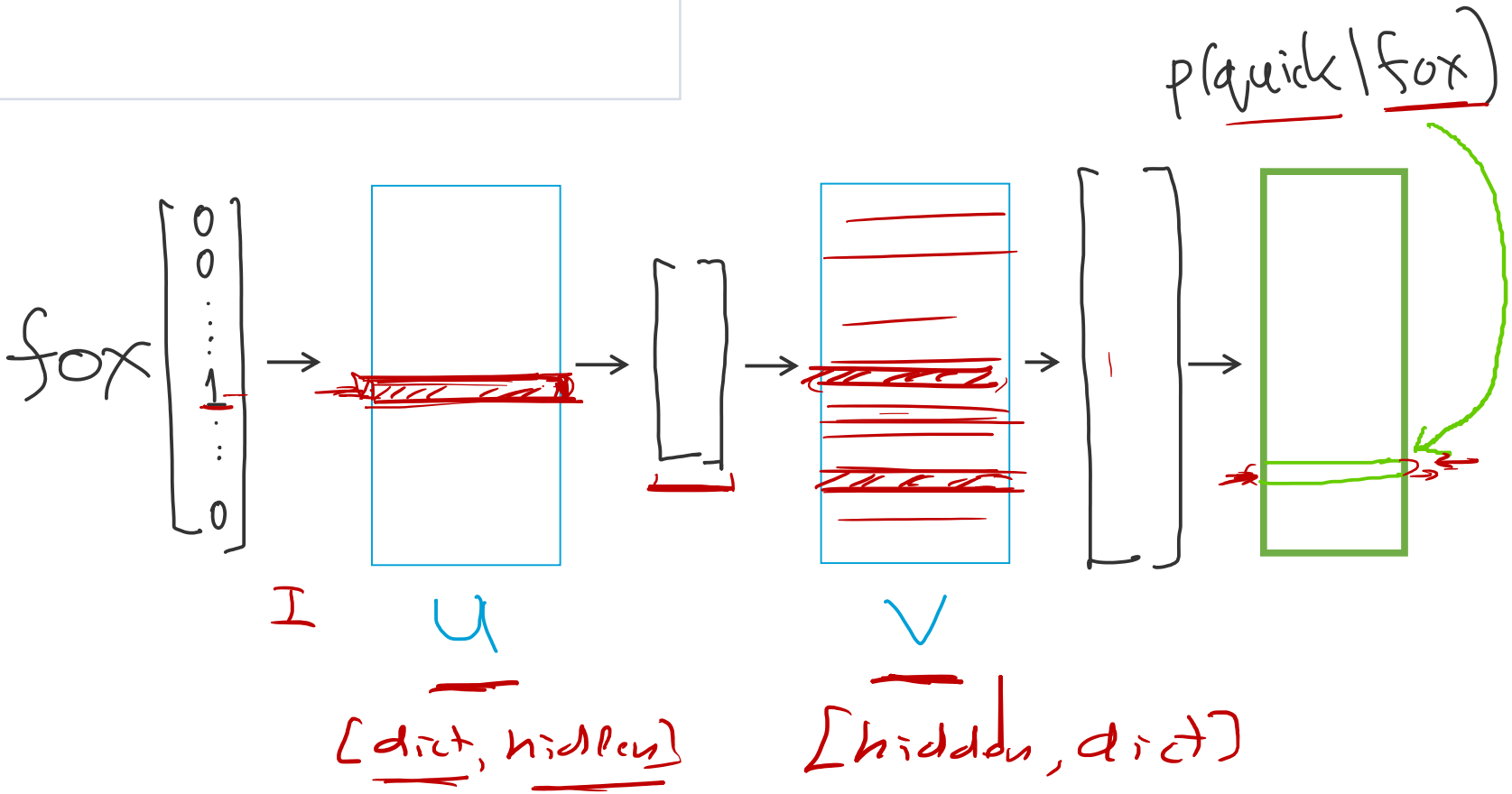
fox

brown

CE

jumps

Word vector



$$p(o|\mathrm{i}) = \frac{e^{u_i \cdot v_o}}{\sum_k e^{u_i \cdot v_k}}$$

$$L = -\sum_j \ln p(o|i)$$

$$= -\sum_j \ln \frac{e^{u_i \cdot v_o}}{\sum_k e^{u_i \cdot v_k}}$$

$$w(fox) = u(fox) + v(fox)$$

fox

[1 0.5 -0.3 ... 0.02]

Word vector

$$p(o|\mathrm{i}) = \frac{\mathrm{e}^{u_i \cdot v_o}}{\sum_k \mathrm{e}^{u_i \cdot v_k}}$$

$$L = -\sum_j \ln p(o|i)$$

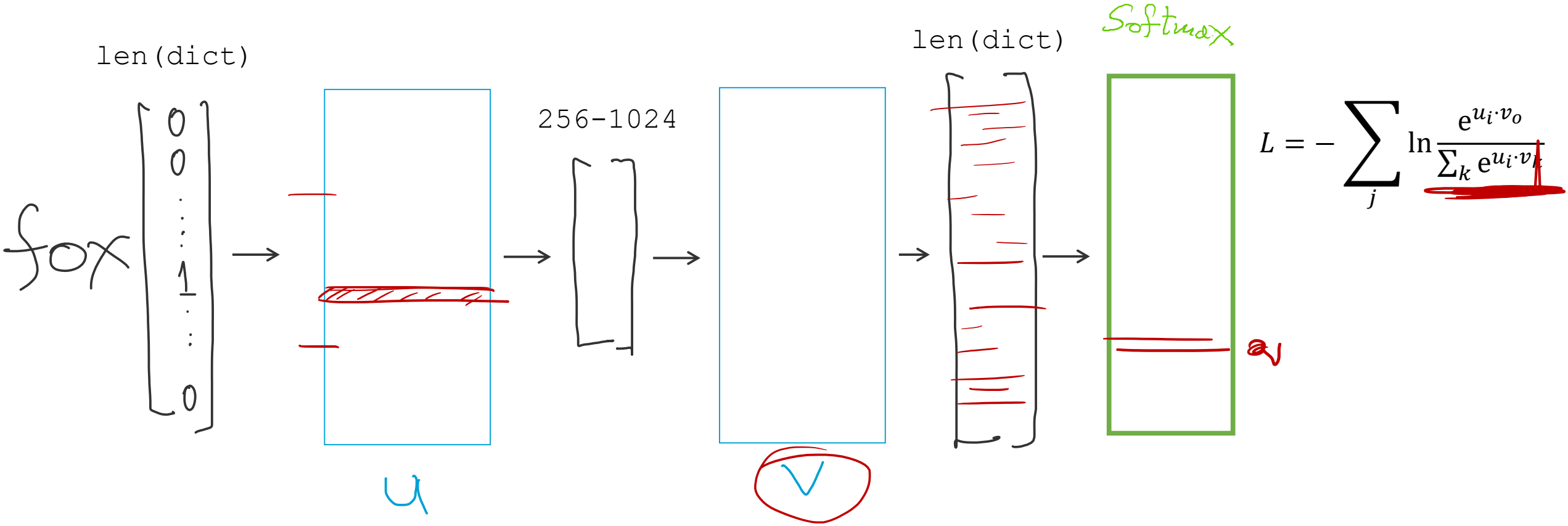
$$= -\sum_j \ln \frac{\mathrm{e}^{u_i \cdot v_o}}{\sum_k \mathrm{e}^{u_i \cdot v_k}}$$

$$w(fox) = u(fox) + v(fox)$$

fox

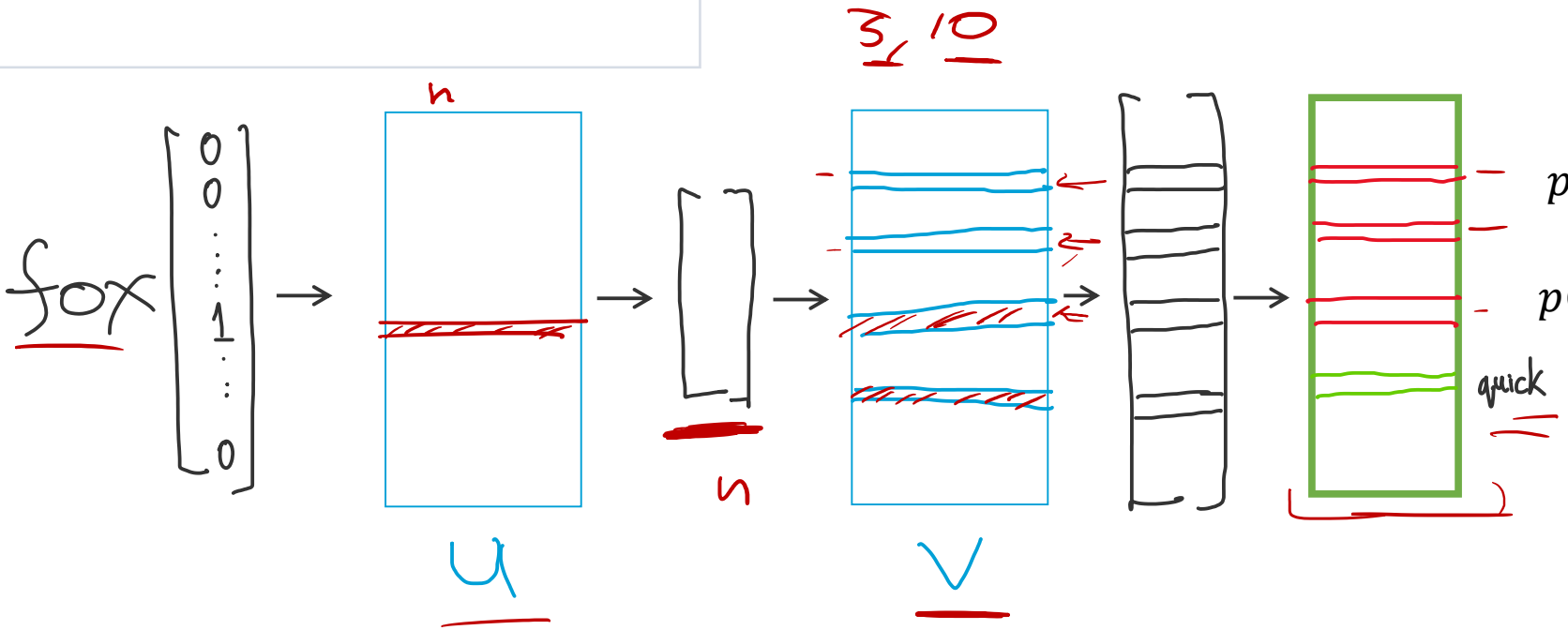
[1 0.5 -0.3 . . . 0.02]

Небольшая проблемка



Negative sampling

broad, jump, quick



$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$p(D = 1|i, o) = \sigma(u_i \cdot v_o) = \frac{1}{1 + e^{-u_i \cdot v_o}}$$

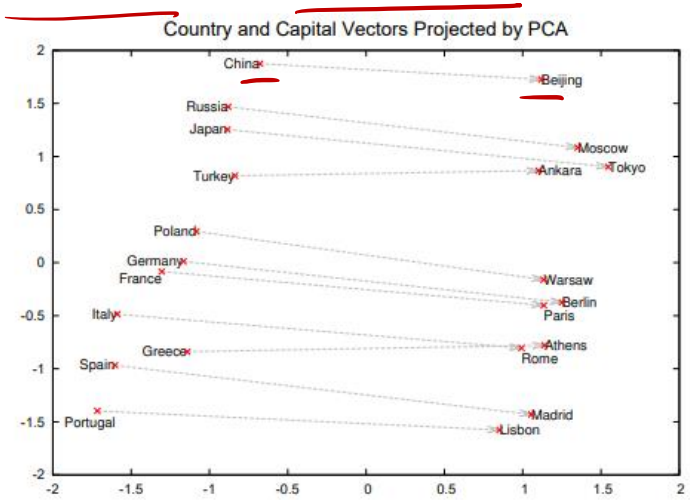
$$p(D = 0|i, o) = 1 - \sigma(u_i \cdot v_o) = \frac{1}{1 + e^{u_i \cdot v_o}}$$

$$L = -\log \sigma(u_i \cdot v_0) - \sum_{k \sim P(\omega)} \log \sigma(-u_i \cdot v_k)$$

$$P(\omega) \sim U(\omega)^{3/4}$$

Свойства word2vec
Аналогии

$$w(Paris) - w(France) + w(Russia) = ? \quad w(Moscow)$$



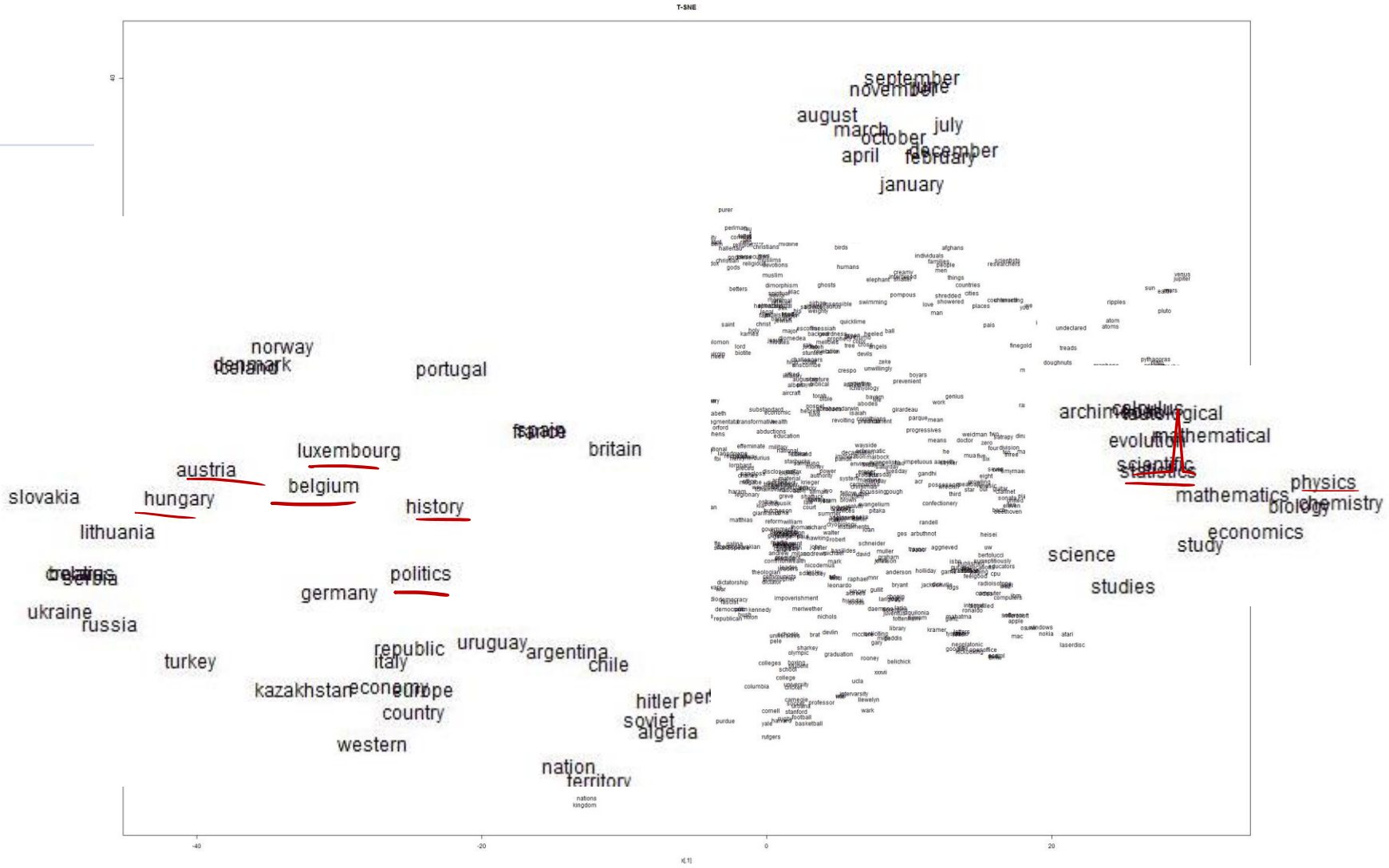
Mikolov'13

king:queen::man:[woman, Attempted abduction, teenager, girl]
house:roof::castle:[dome, bell_tower, spire, crenellations, turrets]
new york times:sulzberger::fox:[Murdoch, Chernin, Bancroft, Ailes]
love:indifference::fear:[apathy, callousness, timidity, helplessness, inaction]
donald trump:republican::barack obama:[Democratic, GOP, Democrats, McCain]
building:architect::software:[programmer, SecurityCenter, WinPcap]

Chris Nicholson on Quora

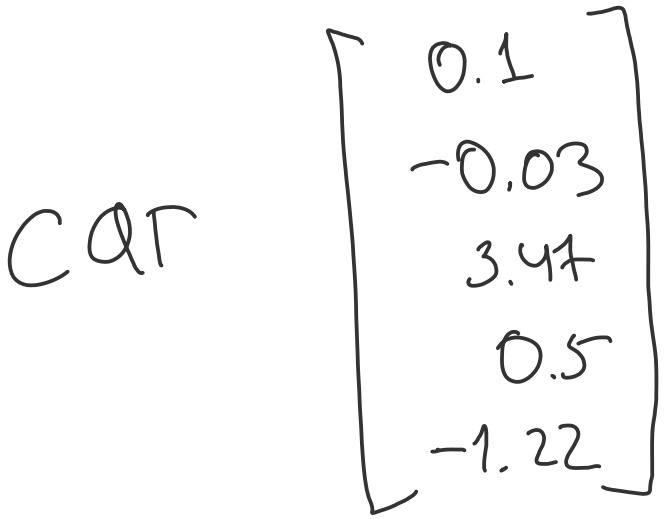
Свойства word2vec
Сходство (similarity)

T-SNE



Source

Ок, а из полезного в хозяйстве?



The film is powerful, accessible and funny
The movie is well done, but slow
Ah yes, and then there's the music

Stanford Sentiment Treebank

$$v(sentence) = \sum_{w} P(word)v(word)$$

fastText от Facebook

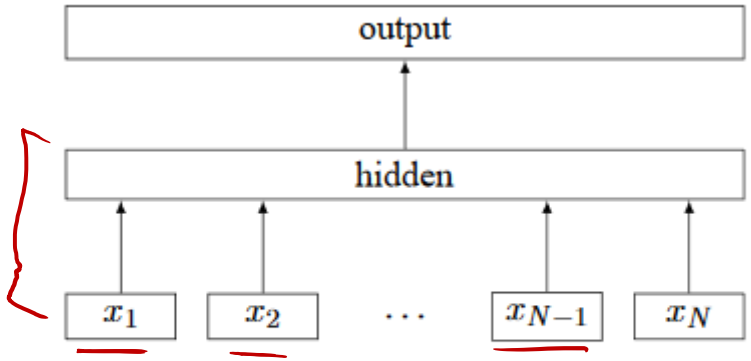


Figure 1: Model architecture of `fastText` for a sentence with N ngram features $x_1, \ldots, x_N$. The features are embedded and averaged to form the hidden variable.

Character N-grams

Each word w is represented as a bag of character n-gram. We add special boundary symbols < and > at the beginning and end of words, allowing to distinguish prefixes and suffixes from other character sequences. We also include the word w itself in the set of its n-grams, to learn a representation for each word (in addition to character n-grams). Taking the word *where* and $n = 3$ as an example, it will be represented by the character n-grams:

`<wh, whe, her, ere, re>`

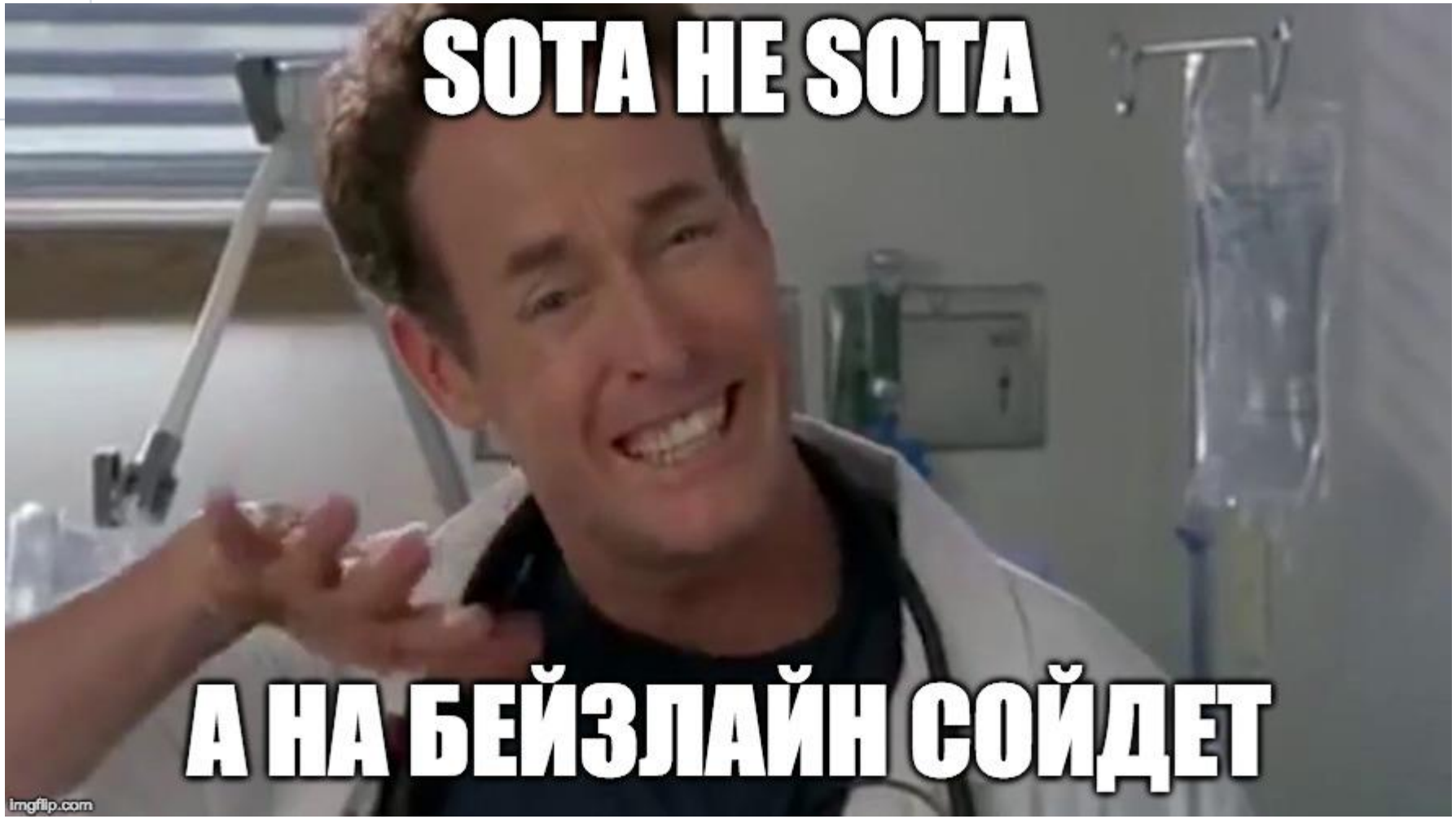
and the special sequence

`<where>`.

усреднить все n-grams

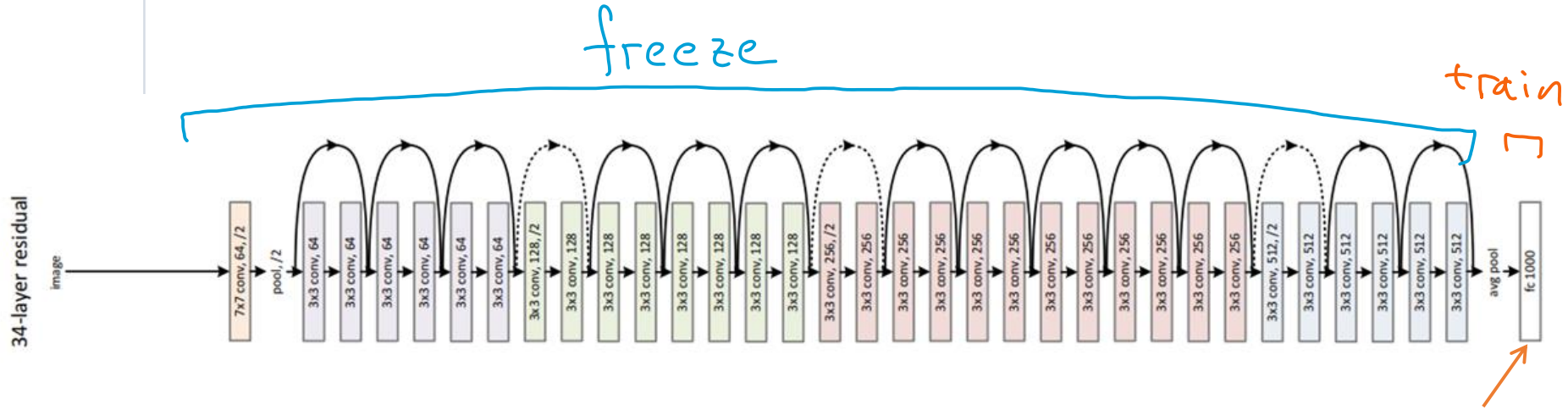
$$s(w, c) = \sum_{g \in \mathcal{G}_w} \mathbf{z}_g^\top \mathbf{v}_c.$$

- Лучше word2vec
- Решает проблему OOV (out of the vocabulary)
- Есть pretrained вектора
- Работает из коробки


SOTA НЕ SOTA
А НА БЕЙЗЛАЙН СОЙДЕТ
imgflip.com


BERT
ELMO
BACK TO THE PRESENT

Pretrained vectors



Линейный классификатор

car → [0.1, -0.03, 3.47, 0.5, -1.22]

```
CLASS  torch.nn.Embedding(
```

```
CLASSMETHOD from_pretrained(
```

Тоже не нужно тренировать самим!

https://code.google.com/archive/p/word2vec/

https://nlp.stanford.edu/projects/glove/

https://github.com/facebookresearch/fastText


ЧТО ЖЕ СЛУЧИТСЯ В СЛЕДУЮЩЕЙ СЕРИИ???
УЗНАЕМ В СЛЕДУЮЩЕЙ СЕРИИ!!!