Введение в принципы модели BERT

Маша Шеянова, masha.shejanova@gmail.com

Пререквизиты

Вспомним Byte Pair Encoding (BPE)

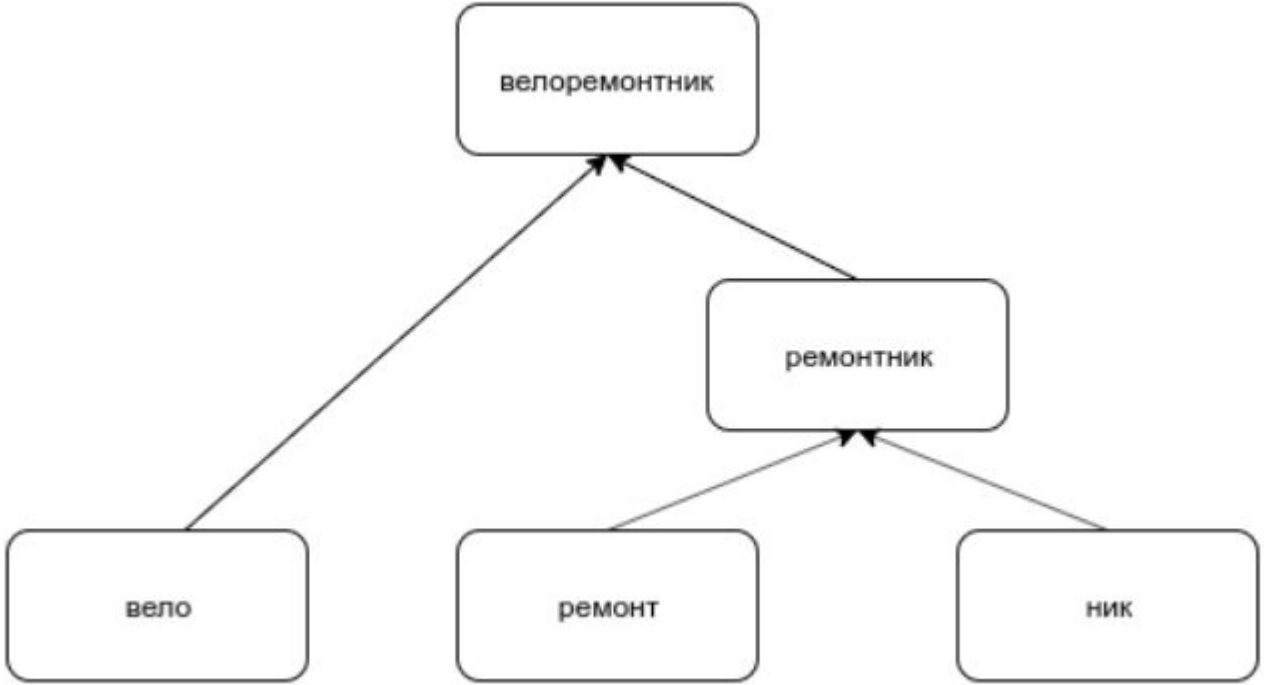
Идея: разбивать текст на меньшие единицы, чем слова, но делать это умно, учитывая **частоту совместной встречаемости**.

Алгоритм:

1) вначале, “юнит” — каждая отдельная буква
2) считаем, какая пара юнитов встречается вместе чаще остальных
3) сливаем такую пару, образуя новый юнит
4) повторяем шаги 2-3, пока не достигнем словаря желаемого размера

В результате — у нас есть юниты разного размера, от слова, до морфемы и, наконец, отдельной буквы. И на такой токенизации можно обучать word2vec.

Пример



Если в обучающем корпусе не было слова *велоремонтник*, то получится *(вело, ремонтник)* или *(вело, ремонт, ник)*.

Чему нас научили seq2seq модели

- **Энкодер** — нейросеть, которая считывает входную последовательность и получает векторное представление входных данных
- **Декодер** — другая нейросеть, которая использует выходы энкодера, чтобы породить новую последовательность
- **Attention** — механизм, которые позволяет понять, насколько для текущего слова важен тот или иной вектор из энкодера (= слово из входной последовательности)
- Attention — это очень хорошо!

Transformers and BERT

Трансформер *(Attention is all you need, 2017)*

(Это очень большая и сложная модель. Если хотите глубоко понять устройство — переходите по ссылкам на слайдах).

В общих чертах:

- как и seq2seq, состоит из энкодера и декодера
- но не использует RNN, **полностью заменив передачу вектора состояния на attention** (и в энкодере это тоже происходит, называется ***self-attention***)
- и энкодер, и декодер многослойные (в оригинальной версии, 6 слоёв), attention применяется на каждом из них

BERT

BERT — это большой энкодер из трансформера, обученный на задаче языкового моделирования.

А именно, его учили:

- предсказывать пропущенные (маскированные) слова в тексте
- угадывать, идёт ли одно предложение за другим

Такие задачи позволяют получить максимально обобщённые для языка параметры, которые можно потом использовать для частных задач, таких как классификация отзывов или новостей.

о применениях BERT

BERT для классификации

Кроме эмбеддингов BPE-токенов, BERT обучает представление добавленного токена [CLS], который “отвечает” за весь текст.

Дальше:

- можно использовать эмбеддинг этого символа как эмбеддинг текста, и сверху добавлять свою модель
- (advanced) можно дообучать его

BERT как контекстные эмбеддинги

BERT использует BPE и в процессе обучения создаёт векторное представление для BPE-токенов.

За счёт self-attention эмбединги каждого слова знают о его контексте.

```
print ("Similarity of 'bank' as in 'bank robber' to 'bank' as in 'bank vault':",  same_bank)
```

```
Similarity of 'bank' as in 'bank robber' to 'bank' as in 'bank vault': 0.9456751
```

```
print ("Similarity of 'bank' as in 'bank robber' to 'bank' as in 'river bank':",  different_ban
k)
```

```
Similarity of 'bank' as in 'bank robber' to 'bank' as in 'river bank': 0.6797334
```

Практика

Использование модели BERT

- для классификации
- как эмбеддинги