Введение в принципы модели BERT

или по верхам о модели, взорвавшей NLP

Маша Шеянова, masha.shejanova@gmail.com

Byte Pair Encodings (BPE)

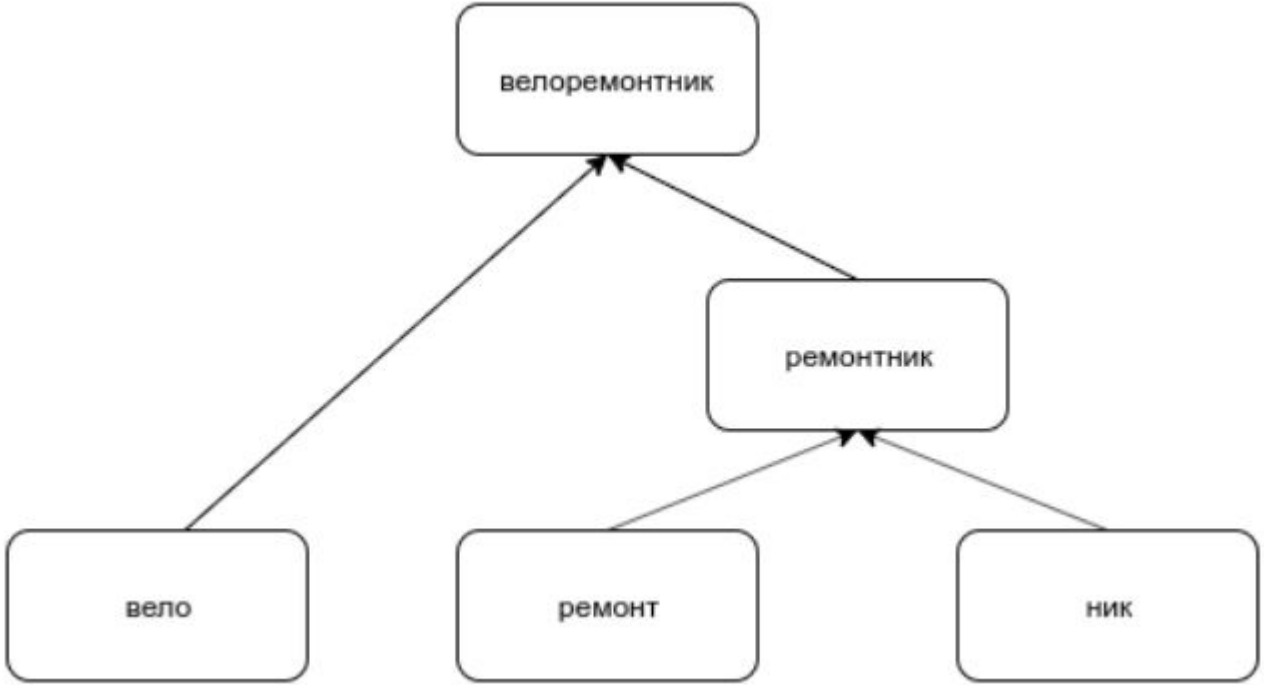
Что такое BPE

Альтернативный способ токенизации для обучения эмбеддингов слов. Чем-то похож на fasttext, но лучше.

Алгоритм:

- вначале у нас столько “токенов”, сколько не-пробельных символов
- пока мы не получили столько токенов, сколько мы хотим в итоге получить
  - находим два самых часто встречаемых друг с другом элемента
  - сливаем, их образуя новый токен
  - повторяем

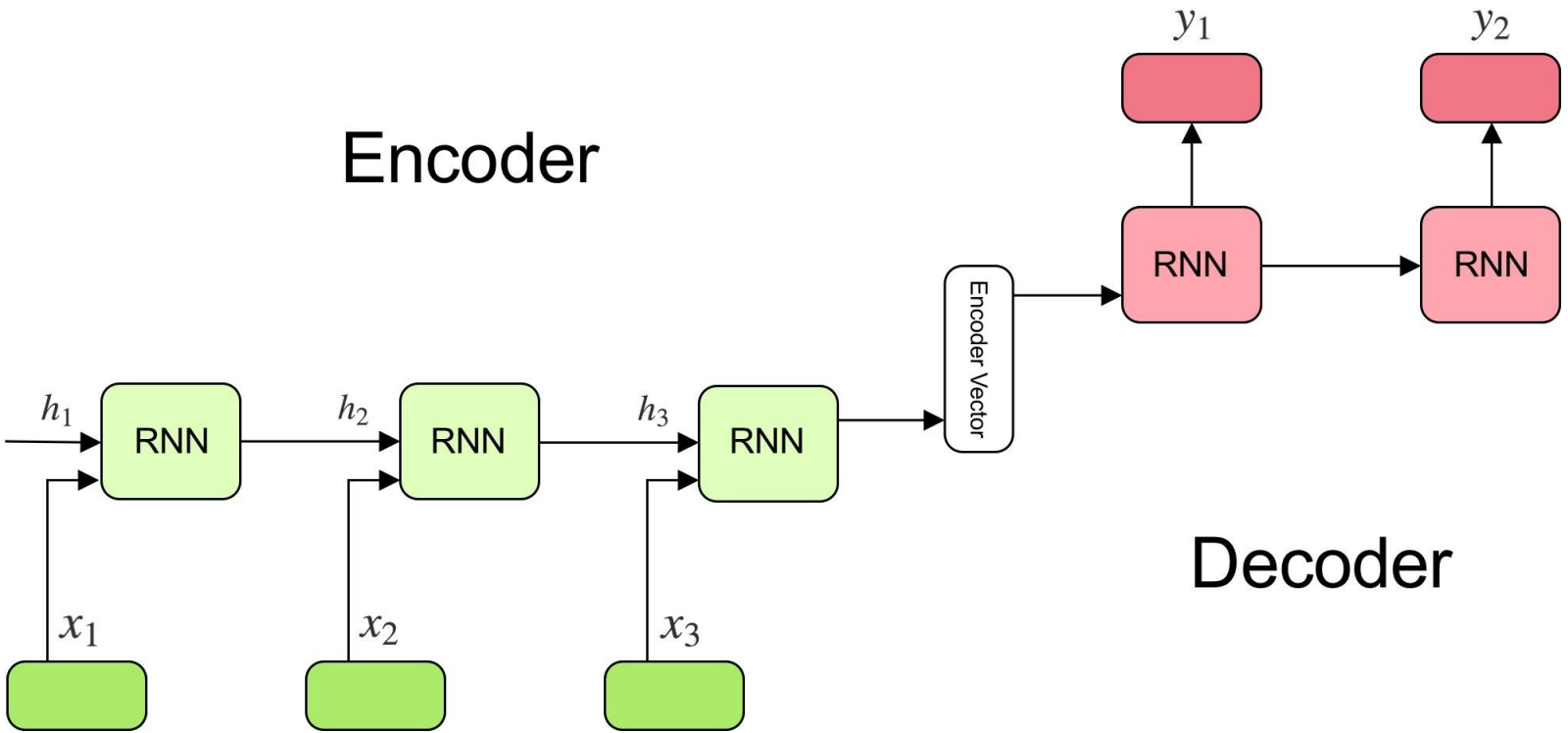
Пример



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

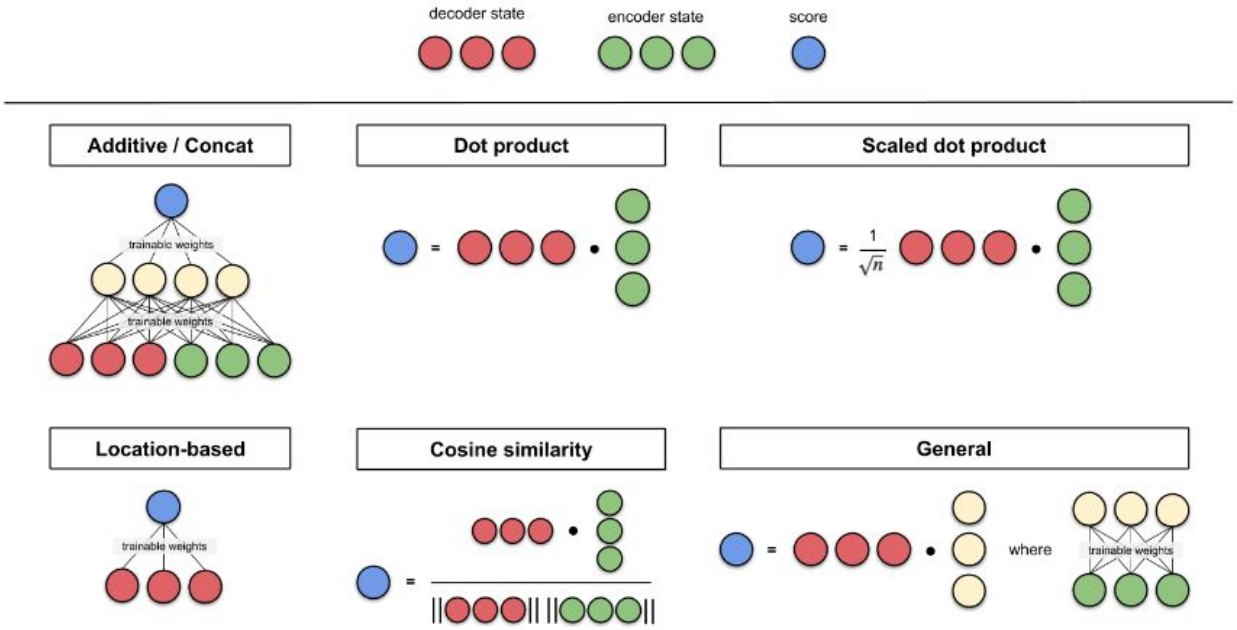
Attention and Transformers

seq2seq (стандартная)



seq2seq + attention

Виды attention (источник)



Виды attention (источник)

Name	Alignment score function	Citation
Content-base attention	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \text{cosine}[\boldsymbol{s}_t, \boldsymbol{h}_i]$	Graves2014
Additive(*)	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \mathbf{v}_a^\top \tanh(\mathbf{W}_a[\boldsymbol{s}_t; \boldsymbol{h}_i])$	Bahdanau2015
Location-Base	$\alpha_{t,i} = \text{softmax}(\mathbf{W}_a \boldsymbol{s}_t)$ Note: This simplifies the softmax alignment to only depend on the target position.	Luong2015
General	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{s}_t^\top \mathbf{W}_a \boldsymbol{h}_i$ where $\mathbf{W}_a$ is a trainable weight matrix in the attention layer.	Luong2015
Dot-Product	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{s}_t^\top \boldsymbol{h}_i$	Luong2015
Scaled Dot-Product(^)	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \frac{\boldsymbol{s}_t^\top \boldsymbol{h}_i}{\sqrt{n}}$ Note: very similar to the dot-product attention except for a scaling factor; where n is the dimension of the source hidden state.	Vaswani2017

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

Это очень большая и сложная модель. Если хотите глубоко понять устройство — переходите по ссылкам на слайде. 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 — это большой энкодер из трансформера, обученный предсказывать пропущенные слова и угадывать, идёт ли одно предложение за другим.

о применениях BERT

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
```

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

Практика

Использование модели BERT

- для классификации
- как эмбеддинги