# Токенизация и практика по RNN

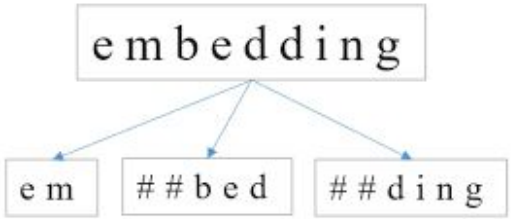
Елена Кантонистова

# Byte Pair Encoding (BPE)

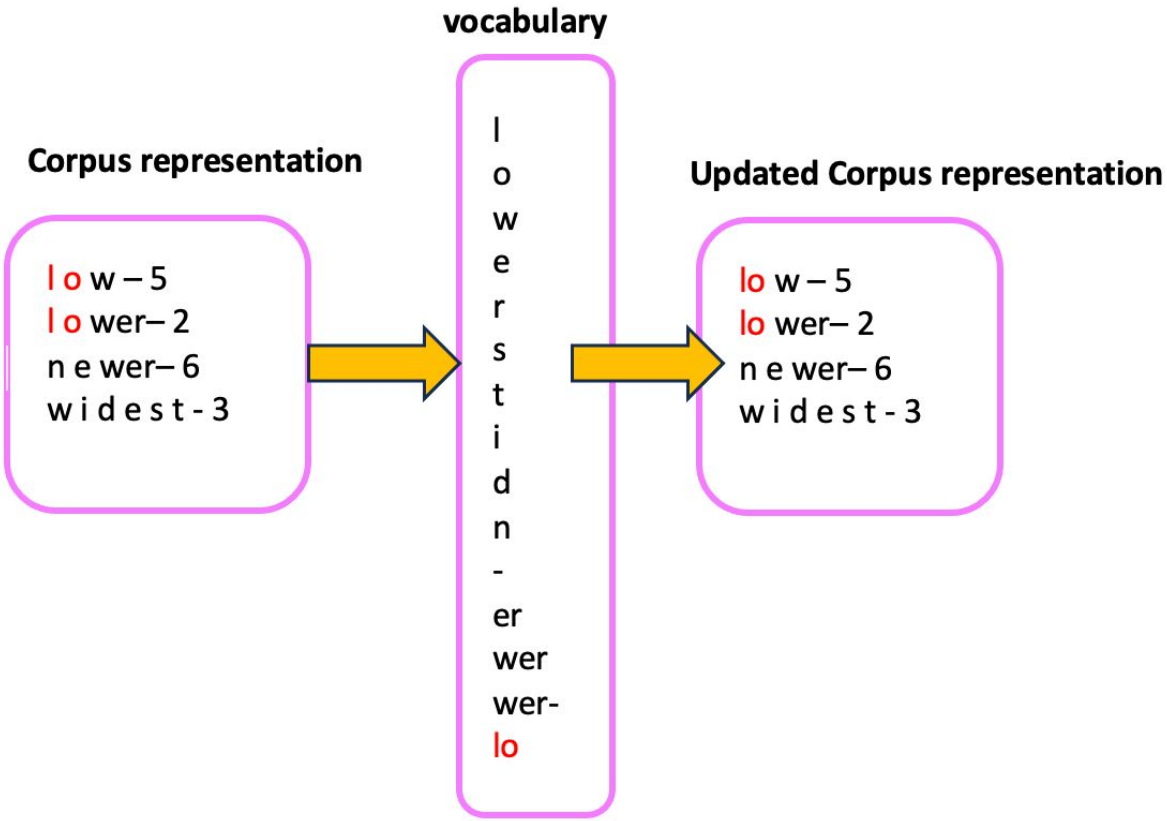
Довольно часто в текстах встречаются новые слова (которых нет в изначальном словаре, на которых модель была обучена), или же новые формы слов, а также редкие токены. Такие слова и токены невозможно закодировать - им будет присвоена метка *UNK (Unknown),* и информацию об этих токенах мы потеряем. Но обычно новое или редкое слово можно разбить на кусочки, каждый из которых уже есть в нашем словаре - будем кодировать эти кусочки, и тогда можно будет кодировать любое слово/токен, и никакая информация о тексте не потеряется. Это мотивация BPE-кодировки.

***Что такое BPE?***

Идея состоит в том, что слова, которые часто встречаются в текстах, мы кодируем как есть (без разбиения), а слова, встречающиеся редко, будут разбиты на кусочки, и отдельные кусочки будут закодированы.



# Byte Pair Encoding (BPE)



# Практика

https://colab.research.google.com/drive/1xfFSviuInMud0Y5ejnj9gtxygUBYQK_G?usp=sharing

# Тестирование по пройденному материалу

Какие проблемы могут возникнуть при использовании посимвольной токенизации?

**Выберите все подходящие ответы из списка**

- [ ] Слишком короткие последовательности
- [ ] Очень большой словарь
- [ ] Очень длинные последовательности
- [ ] Маленький словарь

Чем отличаются подходы One-Hot Encoding (OHE) и Bag of Words (BoW)?

**Выберите один вариант из списка**

- При использовании OHE среди признаков могут появиться отрицательные значения
- Алгоритмы используют различные типы токенизации
- OHE ставит индикатор наличия слова, а BoW считает количество слов
- При использовании BoW вектора схожих по смыслу текстов будут близкими

Какой вывод можно сделать, если у слова высокое значение TF-признака?

**Выберите один вариант из списка**

- Если это слово относится к разряду стоп-слов, то его можно считать важным в рамках своего текста
- Это слово не является важным в рамках своего текста и не является сильным признаком
- Это слово является важным в рамках своего текста и хорошо описывает его особенность
- Если это слово не является стоп-словом, то его можно считать важным для своего текста

При обучении модели Word2Vec нам бы хотелось, чтобы она удовлетворяла некоторым условиям. Выберите из списка те утверждения, которые соответствуют таким условиям.

**Выберите все подходящие ответы из списка**

- [ ] Модель должна улавливать семантику слов
- [ ] Модель должна предполагать объяснимую арифметику векторов
- [ ] Слова с близкими векторами должны быть противоположны по смыслу
- [ ] Близкие по смыслу слова должны иметь близкие векторы

Что является обучаемыми параметрами в модели Word2Vec?

**Выберите все подходящие ответы из списка**

- [ ] Центральные вектора слов
- [ ] Размер словаря
- [ ] Контекстные вектора слов
- [ ] Длина векторов слов

Какой результат наиболее вероятно получится, если после успешного обучения модели Word2Vec на большом корпусе текстов, мы попробуем произвести следующую арифметическую операцию

'Washington' - 'The USA' + 'Norway' = ?

**Выберите один вариант из списка**

- ○ Finland
- ○ New York
- ○ Oslo
- ○ Capital

Какую размерность векторов чаще всего используют при обучении модели Word2Vec?

**Выберите один вариант из списка**

- 350
- 300
- 64
- 256

В каком виде подаются слова в RNN?

Выберите все подходящие варианты ответа.

**Выберите все подходящие ответы из списка**

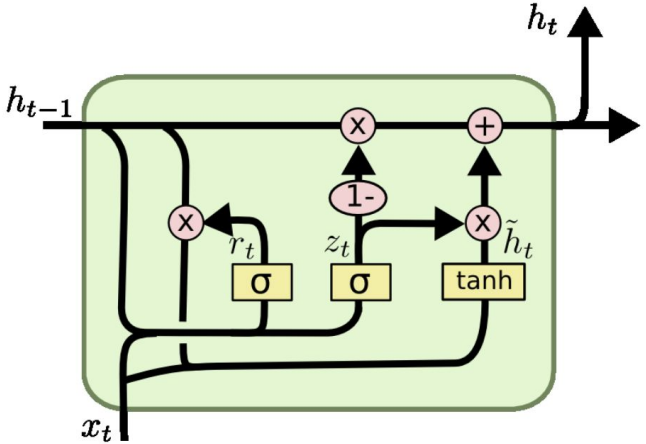
- [ ] На вход подается все предложение, а RNN сама токенизирует и векторизует его
- [ ] Кодируются при помощи word2vec/fasttext и других векторизаторов
- [ ] Кодируются при помощи one-hot encoding
- [ ] В исходном (как токены)

Каким методом обучаются RNN?

**Выберите один вариант из списка**

- Методом производной сложной функции
- Методом BPTT
- Методом градиентного спуска
- Методом обратного распространения ошибки

Как называется нейронная сеть, изображенная на рисунке?




**Выберите один вариант из списка**

- LSTM
- Многослойная RNN
- GRU
- BiLSTM

Какую проблему пытаются решить с помощью двунаправленных LSTM?

**Выберите один вариант из списка**

- ○ Никакую из перечисленных
- ○ Проблему взрыва градиента в chain rule
- ○ Проблему забывания начала длинных текстов
- ○ Проблему зануления вероятностей слов

Из перечисленных выберите все задачи, которые разумно решать seq2seq-архитектурой

**Выберите все подходящие ответы из списка**

- [ ] NER
- [ ] POS-tagging
- [ ] Классификация текстов
- [ ] Машинный перевод
- [ ] Суммаризация текстов