BERT.

Алексей Андреевич Сорокин
Yandex Research,
МГУ, отделение теоретической и прикладной лингвистики.

Школа РАИИ 2021
лекция 5.

Нейронные сети: предобучение на языковом моделировании

- Задача языкового моделирования – предсказание следующего слова в тексте (распределения вероятностей следующего слова).

Я вчера ел вкусную ?

Нейронные сети: предобучение на языковом моделировании

- Задача языкового моделирования – предсказание следующего слова в тексте (распределения вероятностей следующего слова).

Я вчера ел вкусную ?

- Она требует больших знаний о языке:
  - Морфология (если следующее слово существительное, то женского рода).
  - Графика (следующее слово с большой вероятностью кончается на -у).
  - Синтаксис.
  - Семантика (следующее слово “съедобное”).

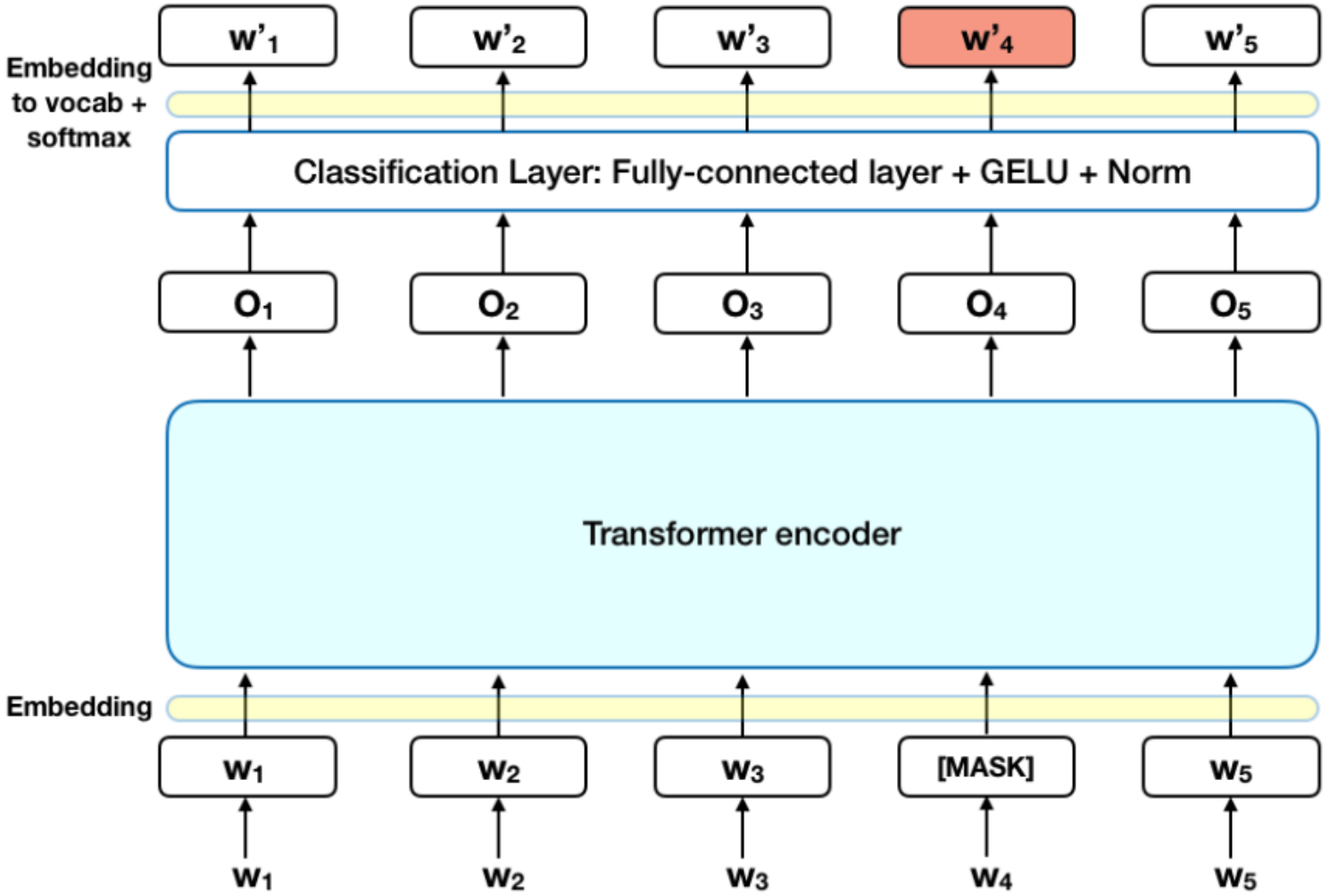
Нейронные сети: предобучение на языковом моделировании

- Задача языкового моделирования – предсказание следующего слова в тексте (распределения вероятностей следующего слова).

Я вчера ел вкусную ?

- Она требует больших знаний о языке:
  - Морфология (если следующее слово существительное, то женского рода).
  - Графика (следующее слово с большой вероятностью кончается на -у).
  - Синтаксис.
  - Семантика (следующее слово “съедобное”).
- Кроме того, для получения этой информации не нужны размеченные данные.

Нейронные сети: BERT



Нейронные сети: BERT

- Глубокая (12 слоёв) архитектура на основе трансформера (механизма внимания).
- Преобразует последовательность индексов токенов в последовательность векторов.
- Дополнительный вектор для токена начала предложения (CLS-токен).

Нейронные сети: BERT

- Глубокая (12 слоёв) архитектура на основе трансформера (механизма внимания).
- Преобразует последовательность индексов токенов в последовательность векторов.
- Дополнительный вектор для токена начала предложения (CLS-токен).
- Предобучается на двух задачах:
  - Восстановлении пропущенного токена.
  - Проверке, идут ли два предложения друг за другом.

Нейронные сети: BERT

- Глубокая (12 слоёв) архитектура на основе трансформера (механизма внимания).
- Преобразует последовательность индексов токенов в последовательность векторов.
- Дополнительный вектор для токена начала предложения (CLS-токен).
- Предобучается на двух задачах:
  - Восстановлении пропущенного токена.
  - Проверке, идут ли два предложения друг за другом.
- Восстановление пропущенных слов – информация на уровне слов.
- Проверка следования предложений – задачи на уровне предложений / пар предложений.

Восстановление пропущенного токена

- Задача восстановления решается для 15% токенов.
  - 80% заменяются на специальный токен $\langle \text{MASK} \rangle$.
  - 10% на произвольное слово.
  - 10% остаются неизменными.
- Это эффективнее, чем предобучать модель слева направо.

BERT: входные данные

- Входное представление:

$$
\begin{aligned}
x_i &= x_i^{token} + x_i^{pos} + x_i^{type}, \\
x_i^{token} &- \text{эмбеддинг текущего токена}, \\
x_i^{pos} &- \text{эмбеддинг позиции } i, \\
x_i^{type} &- \text{эмбеддинг типа токена.}
\end{aligned}
$$

BERT: входные данные

- Входное представление:

$$
\begin{aligned}
x_i &= x_i^{token} + x_i^{pos} + x_i^{type}, \\
x_i^{token} &- \text{эмбеддинг текущего токена}, \\
x_i^{pos} &- \text{эмбеддинг позиции } i, \\
x_i^{type} &- \text{эмбеддинг типа токена}.
\end{aligned}
$$

- Тип токена – $0/1$, нужно в задачах для пары предложений.
- Эмбеддинг позиции - i-ый элемент обучаемой матрицы $max\_length \times d_{hidden}$. Обычно $max\_length = 512, d_{hidden} = 768$.

BERT: токенизация

- BERT рассматривает слова на уровне BPE-токенов (Byte-Pair Encoding).
- BPE-словарь строится по следующим правилам:

BERT: токенизация

- BERT рассматривает слова на уровне BPE-токенов (Byte-Pair Encoding).
- BPE-словарь строится по следующим правилам:
- В начале элементы BPE-словаря – отдельные символы.
- На каждом шаге объединяется самая частая пара:

$$\begin{aligned} \mathrm{t} + \mathrm{h} &\mapsto \mathrm{th}, \\ \mathrm{i} + \mathrm{t} &\mapsto \mathrm{it}, \\ \ldots &\ldots \ldots, \\ \mathrm{th} + \mathrm{e} &\mapsto \mathrm{the}, \\ \ldots &\ldots \ldots \end{aligned}$$

BERT: токенизация

- BERT рассматривает слова на уровне BPE-токенов (Byte-Pair Encoding).
- BPE-словарь строится по следующим правилам:
- В начале элементы BPE-словаря – отдельные символы.
- На каждом шаге объединяется самая частая пара:

$$\begin{aligned} \mathrm{t} + \mathrm{h} &\mapsto \mathrm{th}, \\ \mathrm{i} + \mathrm{t} &\mapsto \mathrm{it}, \\ \dots &\dots \dots, \\ \mathrm{th} + \mathrm{e} &\mapsto \mathrm{the}, \\ \dots &\dots \dots \end{aligned}$$

- Так делается, пока не будет достигнут заранее заданный размер ($\sim$ 30000 в английской модели).

BERT: токенизация

- При токенизации BERT жадно пытается выделить самый длинный токен из словаря, начиная с конца слова.
- Так делается, пока не удастся дойти до начала слова.

BERT: токенизация

- При токенизации BERT жадно пытается выделить самый длинный токен из словаря, начиная с конца слова.
- Так делается, пока не удастся дойти до начала слова.
- При этом различаются токены в начале слова и не в начале:

$$\begin{aligned} \text{playing} &\to \text{play} + \#\#\text{ing}, \\ \text{replay} &\to \text{re} + \#\#\text{play}. \end{aligned}$$

- Наиболее частотные слова состоят из одного токена.

BERT: дообучение

- Задачи, на которых предобучается BERT, это классификация:
  - Языковое моделирование – классификация токенов.
  - Проверка следования – бинарная классификация CLS-токена.
- Непосредственно за классификацию отвечает верхний полносвязный слой с softmax-активацией.

BERT: дообучение

- Задачи, на которых предобучается BERT, это классификация:
  - Языковое моделирование – классификация токенов.
  - Проверка следования – бинарная классификация CLS-токена.
- Непосредственно за классификацию отвечает верхний полносвязный слой с softmax-активацией.
- Архитектура сети практически не изменится, если будет решаться другая задача:
  - Любая задача классификации предложений (или пар предложений).
  - Любая задача классификации отдельных слов.

BERT: дообучение

- При дообучении BERT на новую классификационную задачу заменяется только финальный слой:

$$
\begin{aligned}
\mathrm{p} &= \mathrm{softmax}(W\mathrm{h}), \\
W \in \mathbb{R}^{K*H} &- \text{обучаемая матрица}, \\
K &- \text{число классов}, \\
H &- \text{выходная размерность BERT (обычно 768)}.
\end{aligned}
$$

- Всего с нуля учится только несколько тысяч параметров (матрица W).

BERT: дообучение

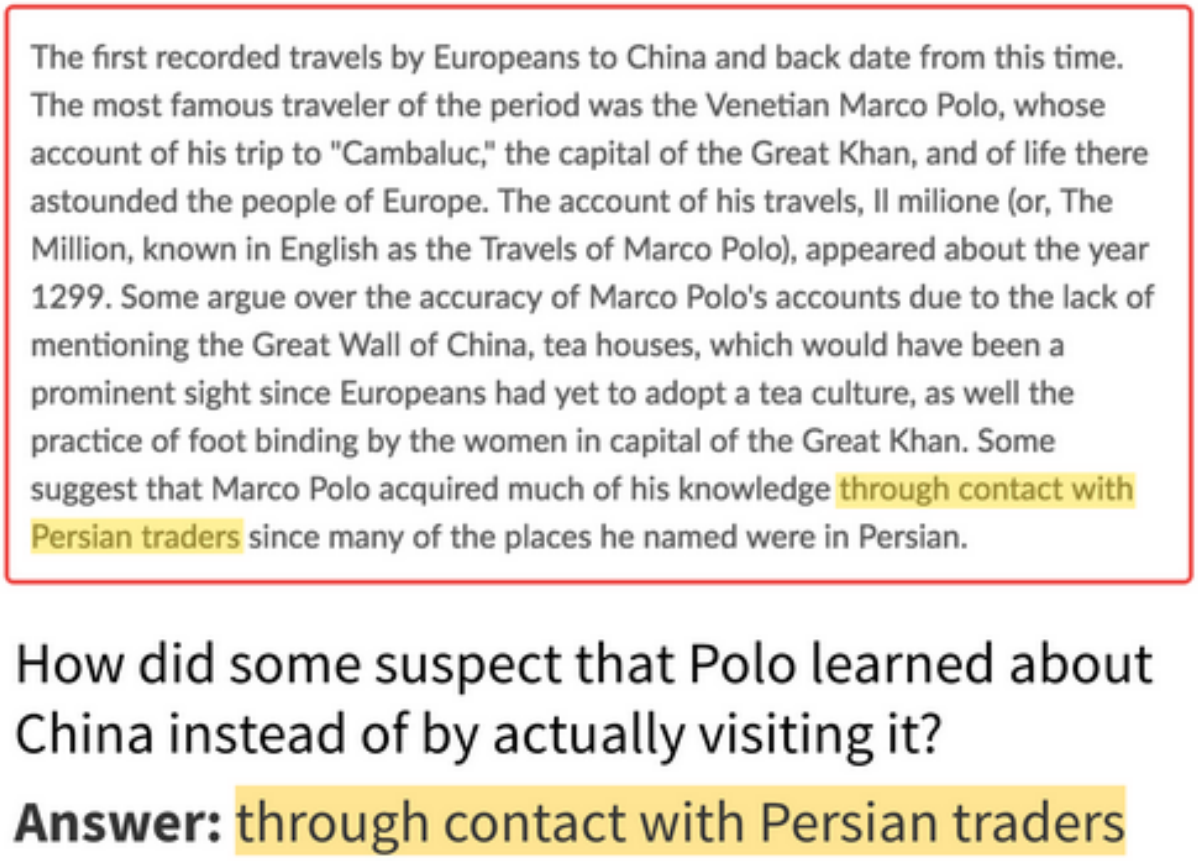
- При дообучении BERT на новую классификационную задачу заменяется только финальный слой:

$$\begin{aligned} \mathrm{p} &= \mathrm{softmax}(W\mathrm{h}), \\ W \in \mathbb{R}^{K*H} &- \text{обучаемая матрица}, \\ K &- \text{число классов}, \\ H &- \text{выходная размерность BERT (обычно 768)}. \end{aligned}$$

- Всего с нуля учится только несколько тысяч параметров (матрица W).
- Это значительно меньше, чем основной энкодер BERT ($\sim 2 * 10^8$ параметров).
- Эта часть сети выучится гораздо быстрее и на небольшом количестве данных.
- При этом веса основной части сети тоже доучивают (обычно они изменятся не так сильно).

BERT: дообучение

- Одна из популярных задач, где BERT сильно улучшил качество – SQuAD:



SQuAD: постановка задачи

- Одна из популярных задач, где BERT сильно улучшил качество – SQuAD.
- В ней требуется выделить в абзаце текста фрагмент, являющийся ответом на вопрос.
- Абзацы взяты из Википедии, вопросы составлены вручную.
- В SQuAD 2.0 появились вопросы, не содержащие ответа.

SQuAD: постановка задачи

- При составлении вопросов рекомендовалось использовать перефразировку, синонимы, гипонимы и гиперонимы:

Passage Segment

...The European Parliament and the Council of the European Union have powers of amendment and veto during the legislative process...

Question

Which governing bodies have veto power?

SQuAD: способы решения

- До появления BERT решали с помощью сопоставления между вопросом и контекстом.
- Например, с помощью разных вариантов внимания.
- Одна из архитектур – BiDAF (Bidirectional Attention Flow):

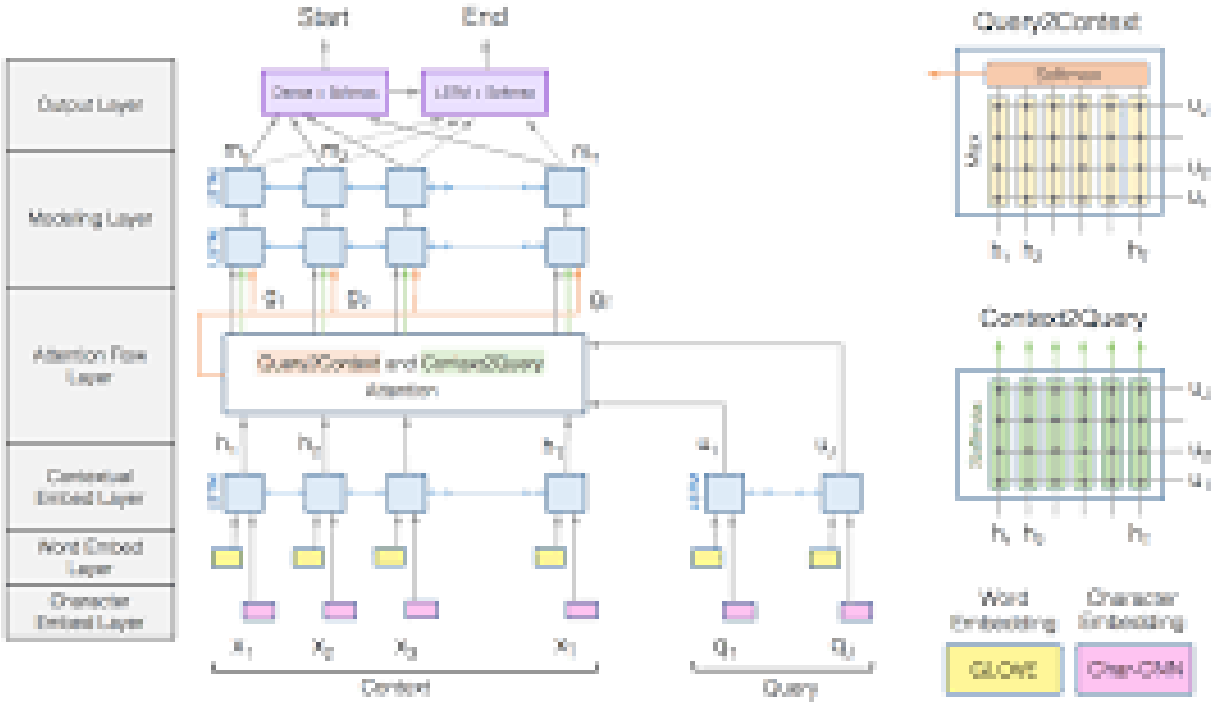
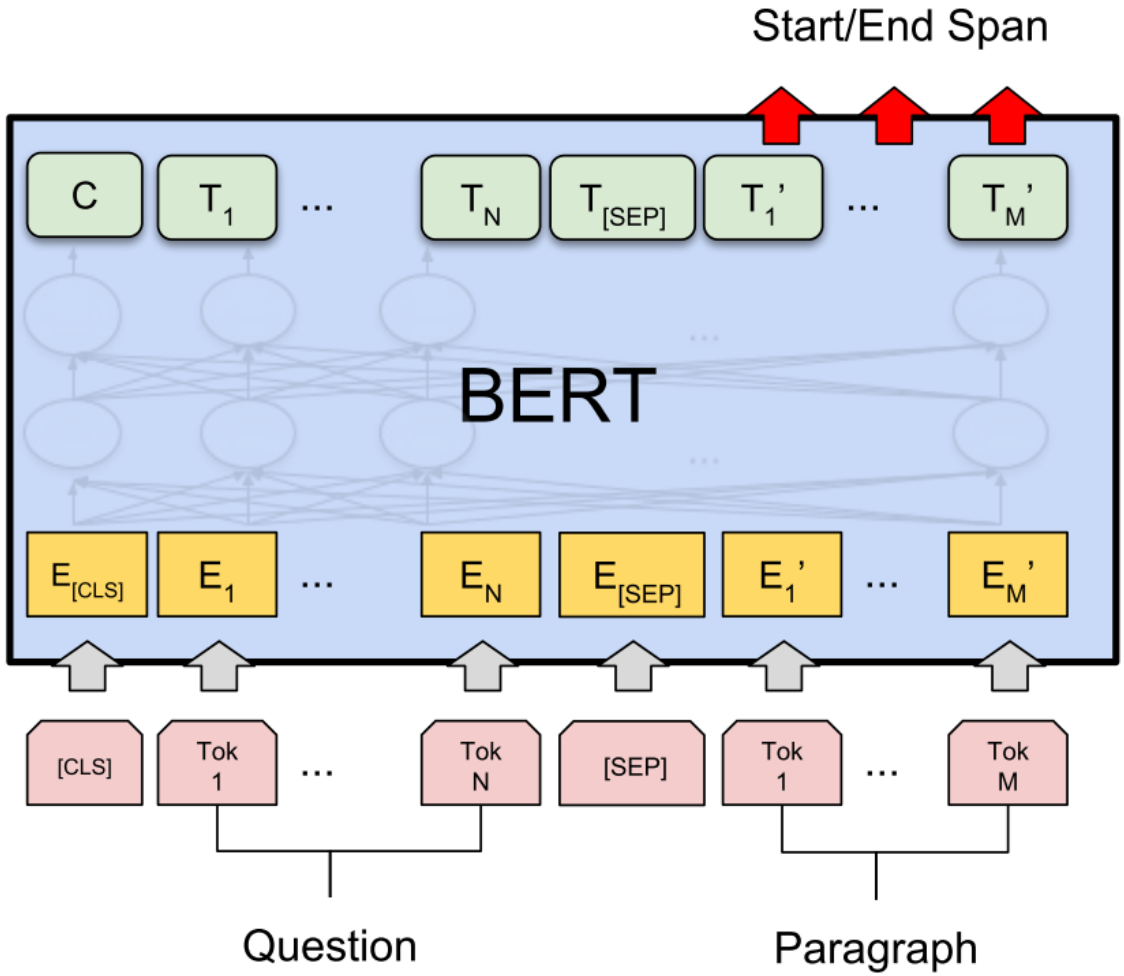


Figure 1: BiDirectional Attention Flow Model (best viewed in color)

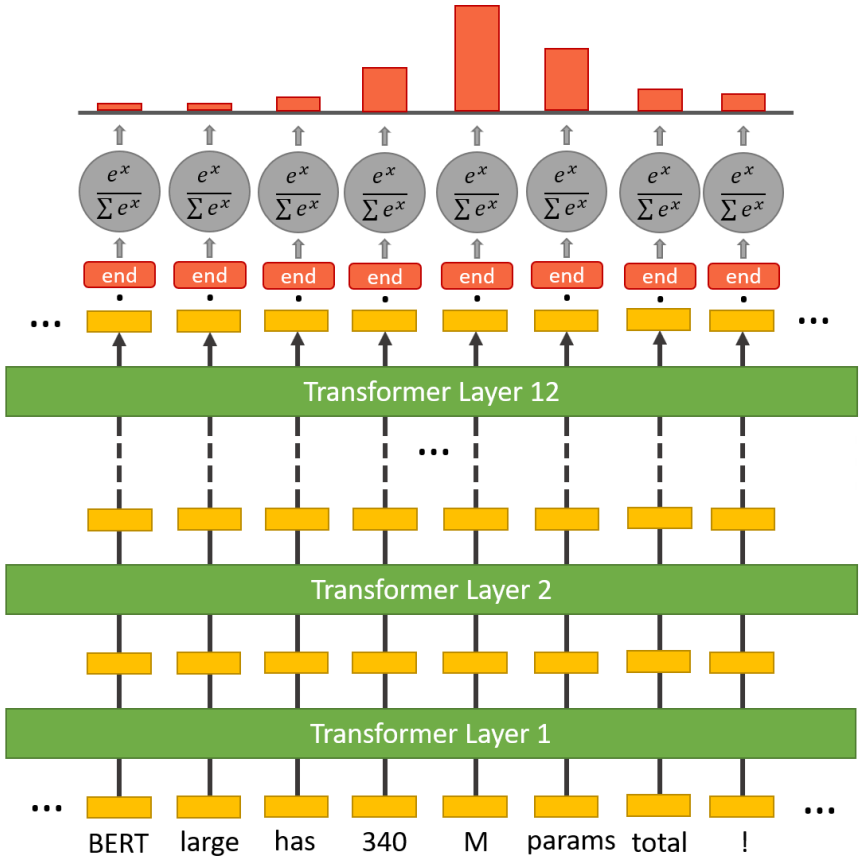
SQuAD: решение с помощью BERT

- Предсказываются позиции начала и конца фрагмента:



SQuAD: решение с помощью BERT

- Для каждой из границ ответ – распределение по токенам:



SQuAD: решение с помощью BERT

- Формальная архитектура (для позиции конца – аналогично):

$$\begin{aligned} [h_1, \ldots, h_n] &= \mathrm{BERT}([x_1, \ldots, x_n]), \\ a_i &= \langle w_S, x_i \rangle, \\ [p_1, \ldots, p_n] &= \mathrm{softmax}([a_1, \ldots, a_n]) \\ w_S &- \text{обучаемый вектор весов.} \end{aligned}$$

- Границы фрагмента выбираются как самая вероятная пара, где начало идёт раньше конца.

Мультиязычная BERT-модель

- Исходный BERT обучался на английской Википедии и корпусе художественной литературы.
- Мультиязычный BERT – на конкатенации 103 основных Википедий.
- Текст из более коротких Википедий сэмплировался с большим весом.

Мультиязычная BERT-модель

- Исходный BERT обучался на английской Википедии и корпусе художественной литературы.
- Мультиязычный BERT – на конкатенации 103 основных Википедий.
- Текст из более коротких Википедий сэмплировался с большим весом.
- Токенизация обучалась на тех же данных с тем же сэмплированием.
- В текст никак не включалась информация о языке.

Мультиязычная BERT-модель

- Исходный BERT обучался на английской Википедии и корпусе художественной литературы.
- Мультиязычный BERT – на конкатенации 103 основных Википедий.
- Текст из более коротких Википедий сэмплировался с большим весом.
- Токенизация обучалась на тех же данных с тем же сэмплированием.
- В текст никак не включалась информация о языке.
- Основное применение мультиязычной модели – настройка моделей для решения задач на конкретном языке.

Мультиязычная BERT-модель: недостатки

- Не все языки одинаково хорошо представлены в мультиязычной модели (точность языковой модели из Rönnquist et al., 2019):

	Mono	Multi
English	**45.92**	33.94
German	**43.93**	28.10
Swedish		22.30
Finnish		14.56
Danish		25.07
Norwegian (Bokmål)		25.21
Norwegian (Nynorsk)		22.28

Мультиязычная BERT-модель: недостатки

- Не все языки одинаково хорошо представлены в мультиязычной модели (точность языковой модели из Rönnquist et al., 2019):

	Mono	Multi
English	**45.92**	33.94
German	**43.93**	28.10
Swedish		22.30
Finnish		14.56
Danish		25.07
Norwegian (Bokmål)		25.21
Norwegian (Nynorsk)		22.28

- Токенизатор тоже может брать частотные фрагменты из другого языка:
  - године
  - року
  - було
  - ##лар.

Обучение BERT для языка

- BERT для любого языка можно было бы обучать так же, как английский.
- Однако обучение с нуля займёт много времени.
- С нуля надо обучать только BPE-токенизацию.

Обучение BERT для языка

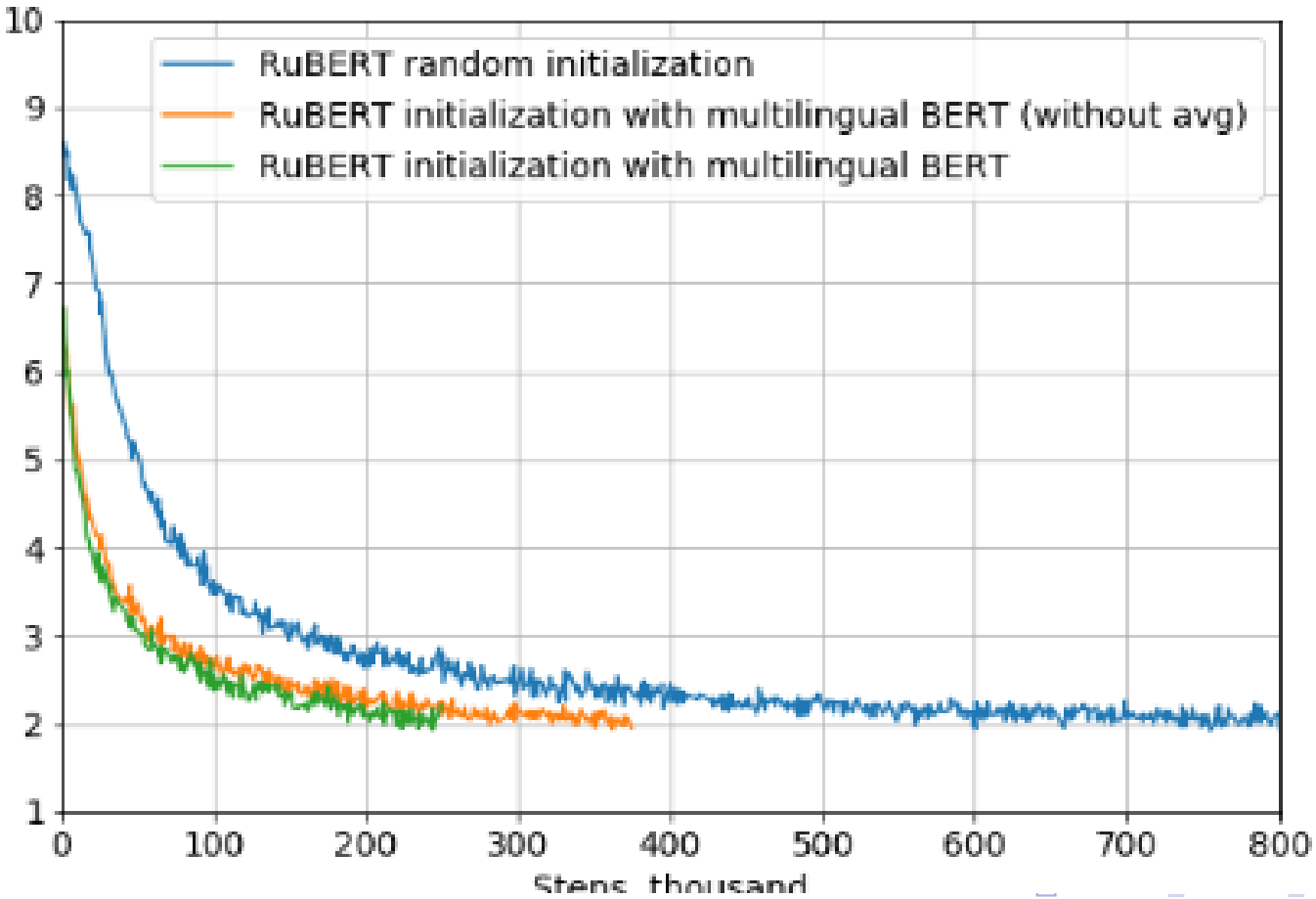
- BERT для любого языка можно было бы обучать так же, как английский.
- Однако обучение с нуля займёт много времени.
- С нуля надо обучать только BPE-токенизацию.
- Можно инициализировать веса слоёв и эмбеддинги сабтокенов весами мультиязычной модели.
- Проблема в том, что набор токенов в словаре поменялся.

Обучение BERT для языка

- BERT для любого языка можно было бы обучать так же, как английский.
- Однако обучение с нуля займёт много времени.
- С нуля надо обучать только BPE-токенизацию.
- Можно инициализировать веса слоёв и эмбеддинги сабтокенов весами мультиязычной модели.
- Проблема в том, что набор токенов в словаре поменялся.
- Решение: эмбеддинг токена инициализируется усреднением его мультиязычных эмбеддингов.

Обучение BERT для русского языка

- За счёт инициализации обучение существенно ускоряется (Kuratov and Arkhipov, 2019):



Обучение BERT для русского языка

- Улучшается качество на парафразе и ответах на вопросы:

model	F-1	Accuracy
Neural networks [11]	79.82	76.65
Classifier + linguistic features [11]	81.10	77.39
Machine Translation + Semantic similarity [6]	78.51	81.41
BERT multilingual	85.48 ± 0.19	81.66 ± 0.38
RuBERT	$\mathbf{87.73} \pm 0.26$	$\mathbf{84.99} \pm 0.35$

Table 1: ParaPhraser. We compare BERT based models with models in non-standard run setting, when all resources were allowed.

model	F-1 (dev)	EM (dev)
R-Net from DeepPavlov [2]	80.04	60.62
BERT multilingual	83.39 ± 0.08	64.35 ± 0.39
RuBERT	$\mathbf{84.60} \pm 0.11$	$\mathbf{66.30} \pm 0.24$

Обучение BERT для нескольких языков

- Можно обучать BERT и для нескольких родственных языков одновременно:

Model	Span F_1	RPM	REM	SM
Bi-LSTM-CRF (Lample et al., 2016)	75.8	73.9	72.1	72.3
Multilingual BERT [5]	79.6	77.8	76.1	77.2
Multilingual BERT-CRF	81.4	80.9	79.2	79.6
Slavic BERT	83.5	83.8	82.0	82.2
Slavic BERT-CRF	87.9	85.7 (90.9)	84.3 (86.4)	84.1 (85.7)

Table 1: Metrics for BSNLP on validation set (Asia Bibi documents). Metrics on the test set are in the brackets.

Обучение BERT для нескольких языков

- Побочный недостаток дообучения BERT: падает качество на родственных языках (пример для задачи морфосинтаксического анализа, Sorokin, 2019)

Training data	BERT	Tag	Tag sent	LAS	Sent LAS	UAS	Sent UAS
be	multlingual	85,09	10,29	76,34	14,71	83,72	17,65
be	Russian	80,75	4,41	45,66	1,47	57,45	4,41
be+ru+uk	multlingual	**88,57**	**19,12**	**84,8**	**16,18**	**90,74**	**33,82**
be+ru+uk	Russian	83,79	7,35	59,3	1,47	68,74	4,41

- То есть модель очень сильно переобучается под конкретный язык.