Методы предобработки текстовых данных для ускорения обучения языковых моделей

Сурков Максим Константинович

Научный руководитель: Ямщиков Иван Павлович

Санкт-Петербургская школа физико-математических и компьютерных наук
НИУ ВШЭ СПБ

17 марта 2021 г.

Обработка естественного языка в реальной жизни

- социальные сети
- электронная почта
- службы доставки
- голосовые помощники
- переводчики
- чат боты




Задачи обработки естественного языка

1. классификация последовательностей
   - спам
   - грубая речь[1]
2. генерация выходной последовательности из исходной
   - машинный перевод
   - ответы на вопросы
3. выделение информации из последовательностей
   - выделение именованных сущностей[2]

[1] G. H. Paetzold et al., SemEval'19 Task 5: Hate Speech Identification with RNN.

[2] Vikas Yadav et al., SemEval'19 Task 12: Deep-Affix Named Entity Recognition of Geolocation Entities. ACL'19

Современные методы решения задач обработки естественного языка

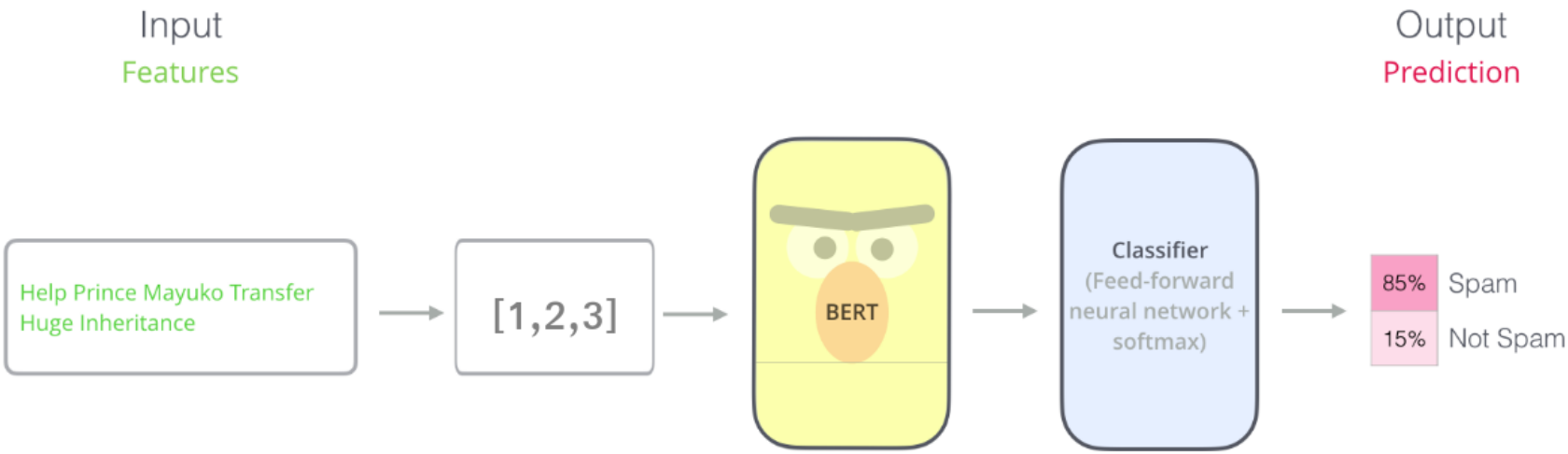
1. Механизм внимания[1]
2. **BERT** (Google)[2]
3. GPT-3 (OpenAI)[3]

[1] Ashish Vaswani et al., Attention Is All You Need, 2017

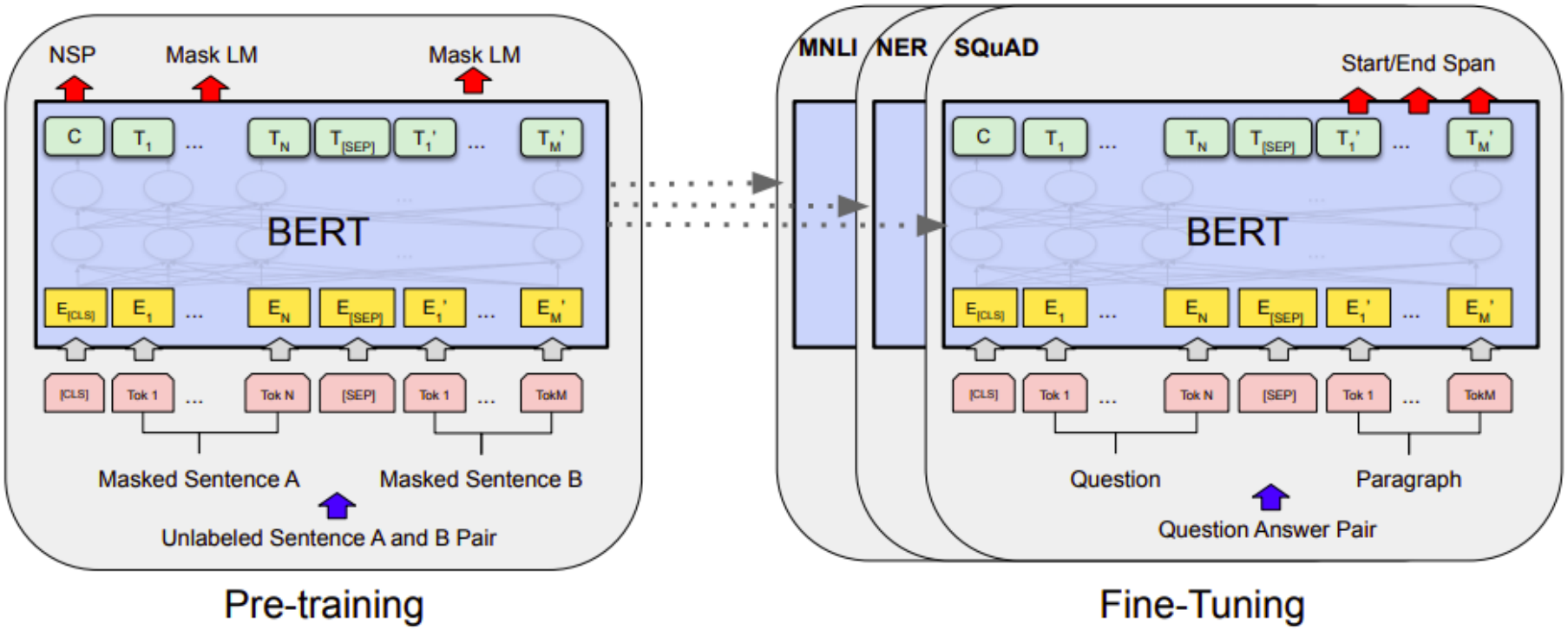
[2] Jacob Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019

[3] Tom B. Brown et al., Language Models are Few-Shot Learners, 2020

BERT. Использование



BERT. Обучение



BERT. Требуемые ресурсы

- количество параметров: $110M - 340M$
- время на предобучение: от 2-4 дней до 1-2 недель[1]
  - мировой рекорд: 47 минут на **1472** V100 GPU[2]
- время на дообучение: 1-2 дня
- размеры данных:

Датасет	Размер
Wikipedia	3-600M
HND	600k-2M
s140	1.6M
IWSLT	200-230k
QQP	364k
MNLI	393k

[1] При использовании 1x-4x GPU Nvidia Tesla V100 32Gb

[2] https://developer.nvidia.com/blog/training-bert-with-gpus

BERT. Существующие методы оптимизации

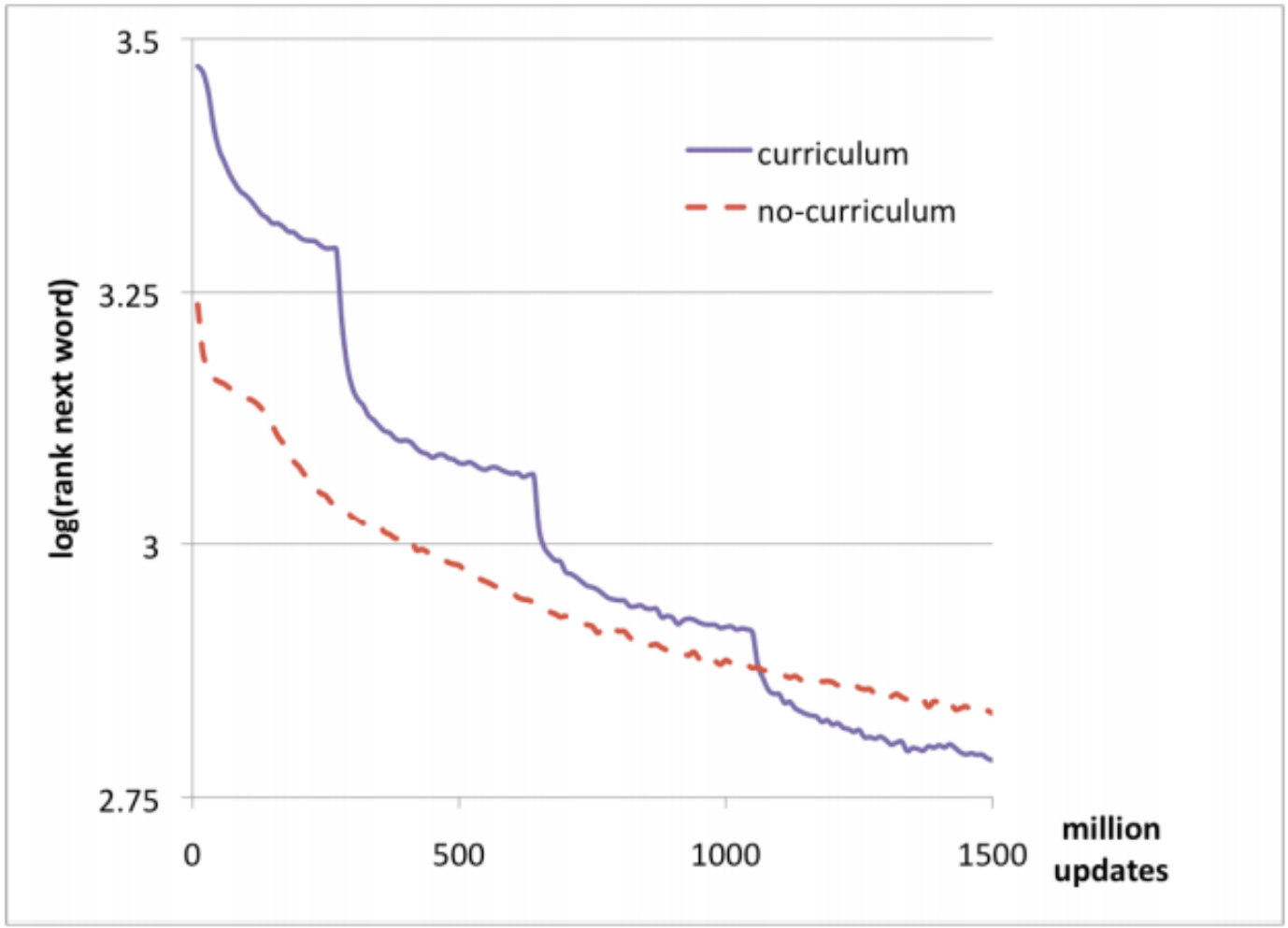
- квантизация[1]
- дистилляция[2]
- прунинг[3]

[1] Sheng Shen et al., Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT, 2019

[2] Victor Sanh et al., DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2020

[3] Hassan Sajjad et al., Poor Man's BERT: Smaller and Faster Transformer Models, 2020

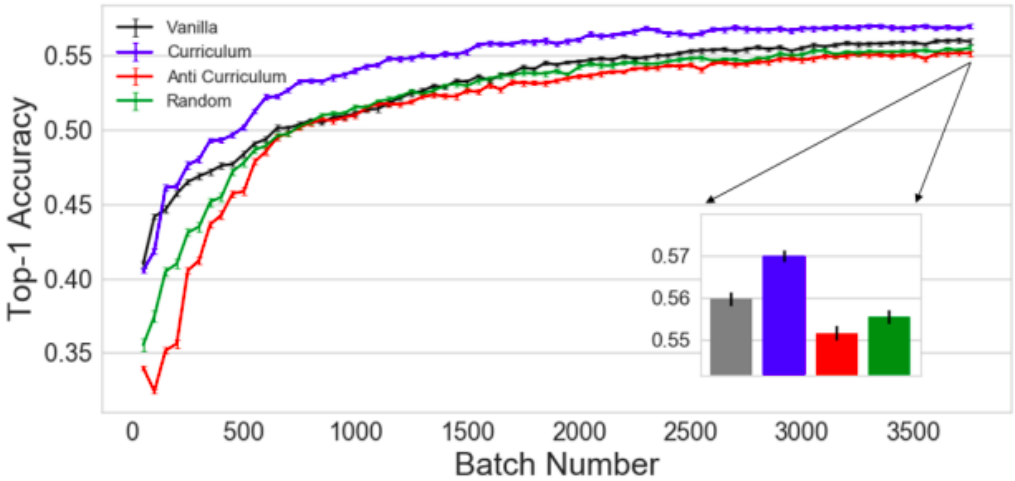
Обучение с расписанием. Начало



Y. Bengio et al., Curriculum learning, 2009

Обучение с расписанием. Применение

- компьютерное зрение[1]



- обучение с подкреплением[2]
- глубокое обучение[3]

---

[1] Guy Hacohen, Daphna Weinshall, On The Power of Curriculum Learning in Training Deep Networks, 2019

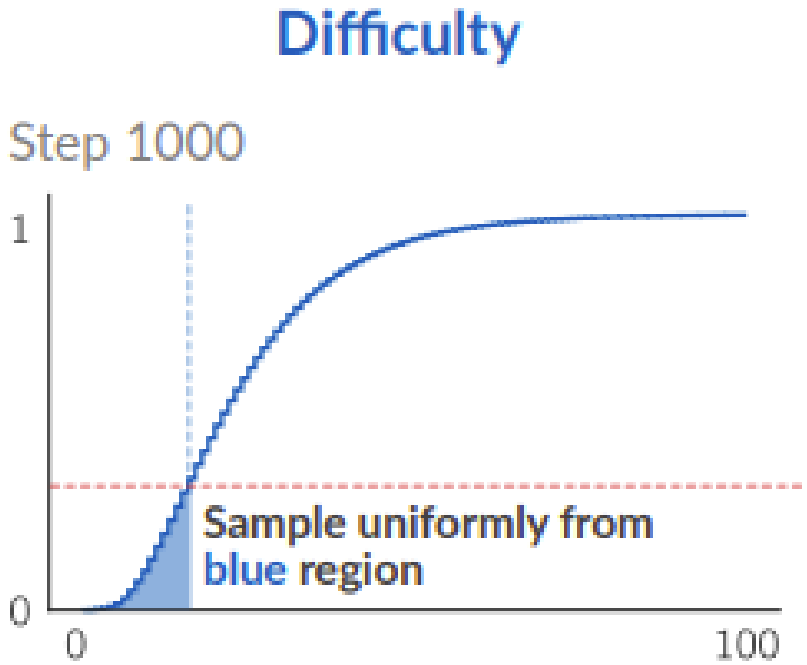
[2] Sanmit Narvekar et al., Curriculum Learning for Reinforcement Learning Domains: A Framework and Survey, 2020

[3] Mermer et al., Scalable Curriculum Learning for Artificial Neural Networks, 2017

Обучение с расписанием в обработке языка

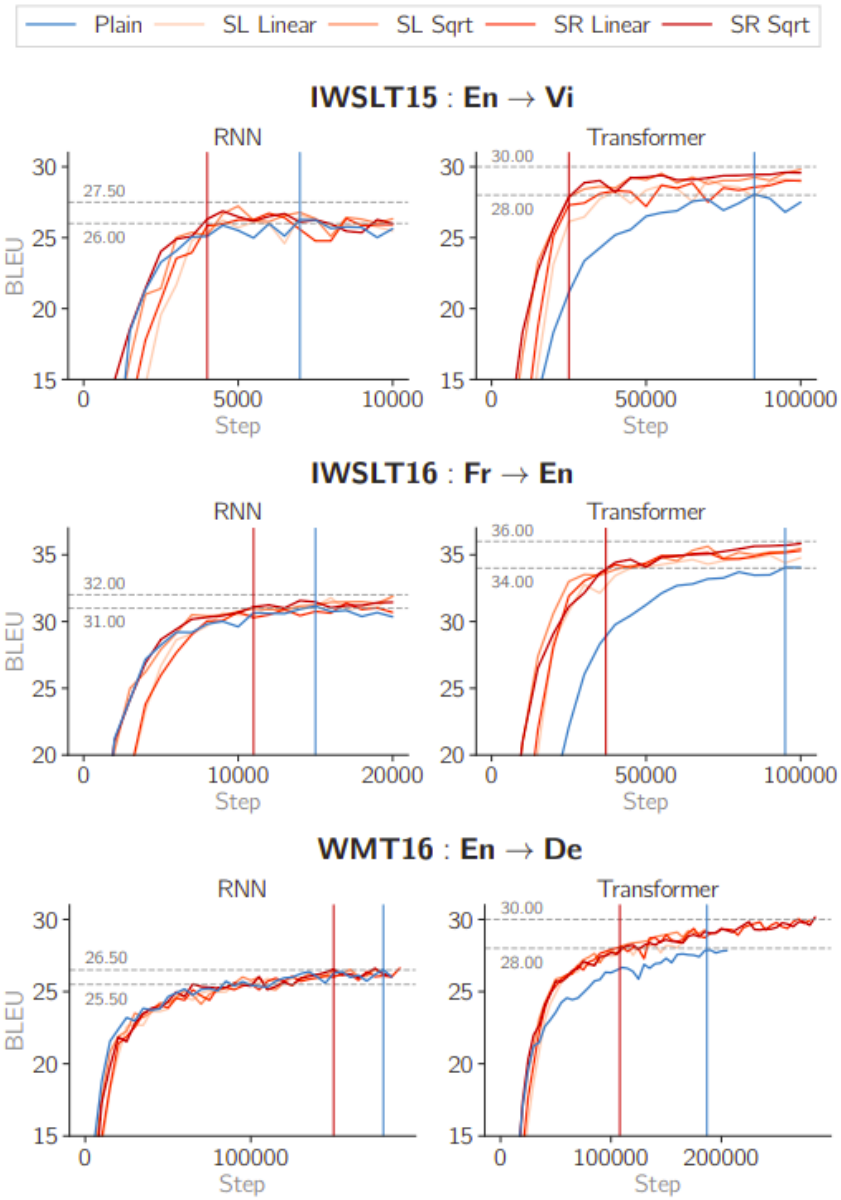
- Задача: машинный первод
- Модель: BERT, LSTM
- Датасеты: IWSLT'15, IWSLT'16, WMT'16
- Алгоритм:

1. сортируем тексты по сложности (длина, логарифм веротности правдоподобия)
2. в течение T шагов (рассмотрим шаг t)
   - считаем $c(t) \in [0, 1]$
   - строим батч из $c(t)$ **первых** текстов корпуса
   - шаг обучения



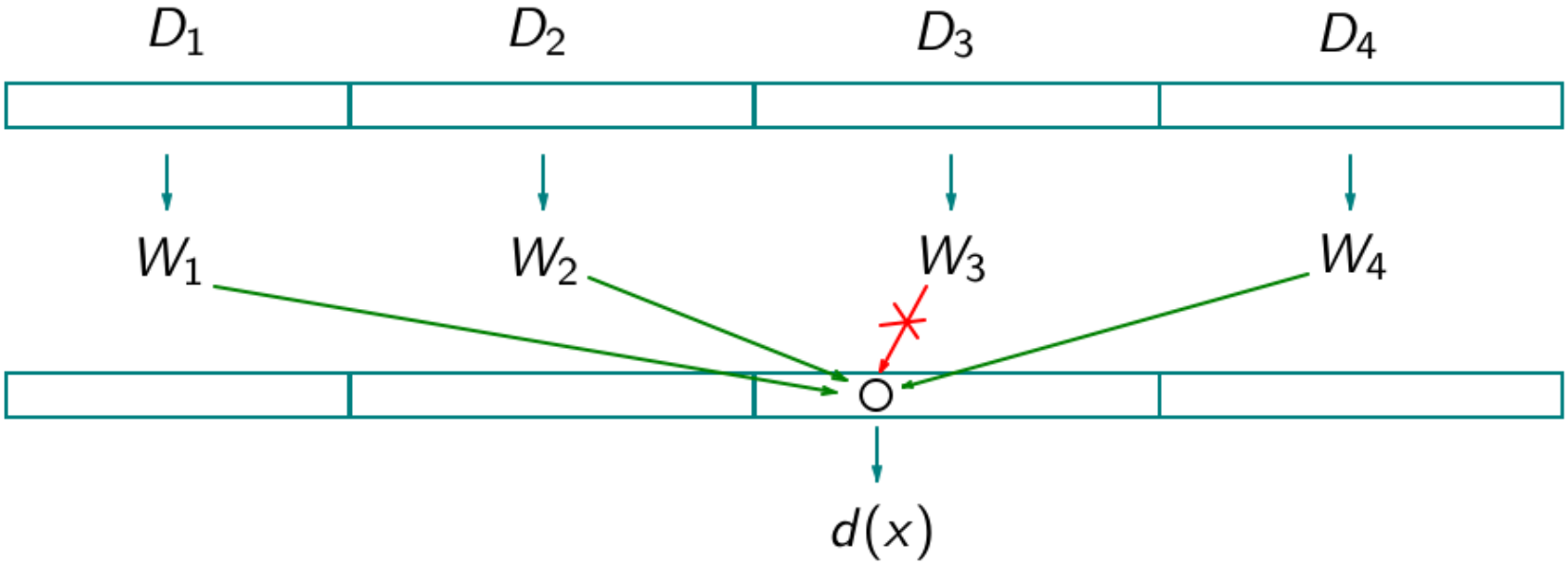
E. A. Platanios et al., Competence-based Curriculum Learning for Neural Machine Translation, ACL'19

Обучение с расписанием в обработке языка



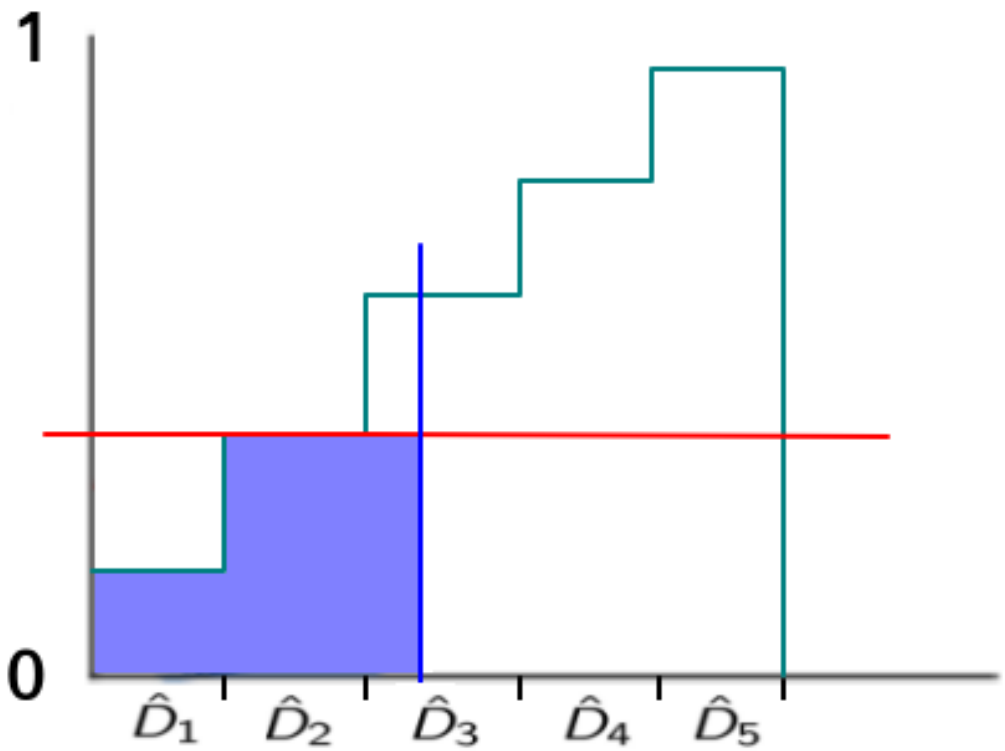
Обучение с расписанием в обработке языка

- Задача: классификация
- BERT
- Датасеты: SQuAD 2.0, NewsQA, GLUE
- Алгоритм: в течение T шагов



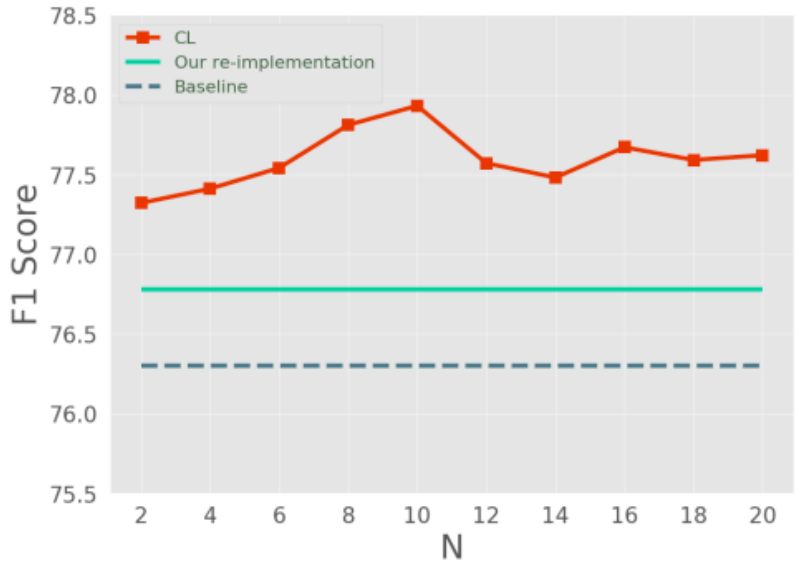
Benfeng Xu et al., Curriculum Learning for Natural Language Understanding, ACL'20

Обучение с расписанием в обработке языка



Обучение с расписанием в обработке языка

	MNLI-m	QNLI	QQP	RTE	SST-2	MRPC	CoLA	STS-B	Avg
results on dev									
BERT Large	**86.6**	92.3	91.3	70.4	93.2	88.0	60.6	90.0	84.1
BERT Large*	**86.6**	92.5	91.5	74.4	93.8	91.7	63.5	90.2	85.5
BERT Large+CL	**86.6**	**92.8**	**91.8**	**76.2**	**94.2**	**91.9**	**66.8**	**90.6**	**86.4**
results on test									
BERT Large	**86.7**	91.1	89.3	70.1	**94.9**	89.3	60.5	87.6	83.7
BERT Large*	86.3	92.2	**89.5**	70.2	94.4	89.3	60.5	87.3	83.7
BERT Large+CL	**86.7**	**92.5**	**89.5**	**70.7**	94.6	**89.6**	**61.5**	**87.8**	**84.1**



Обучение с расписанием в обработке языка. Направления для исследований

- Много важных задач обработки естественного языка с **большими корпусами тренировочных данных**
- Решаются с помощью **тяжелых** моделей, которые **долго** учатся
- Не исследованы метрики оценки сложности текста (длина - текущий предел)
- Эксперименты проведены только на определенных задачах
  - ACL'19 - только задача машинного перевода
  - ACL'20 - только задача классификации[1]
- Не исследовано влияние обучения с расписанием на этапе предобучения

[1] Не совсем честное обучение с расписанием; Не ускоряет; Требует еще больших ресурсов

Цели и задачи

Цель: ускорить обучение языковой модели BERT с помощью обучения с расписанием за счет метрики оценки сложности текстовых данных на задачах предобучения, классификации и машинного перевода

Задачи:

1. Найти эффективные[1] метрики оценки сложности текста
2. Реализовать механизм подсчета найденных метрик на больших датасетах
3. Исследовать влияние найденных метрик на скорость обучения языковой модели BERT
4. Сравнить найденные метрики с существующими метриками оценки сложности текста

[1] с точки зрения сокрости обучения модели