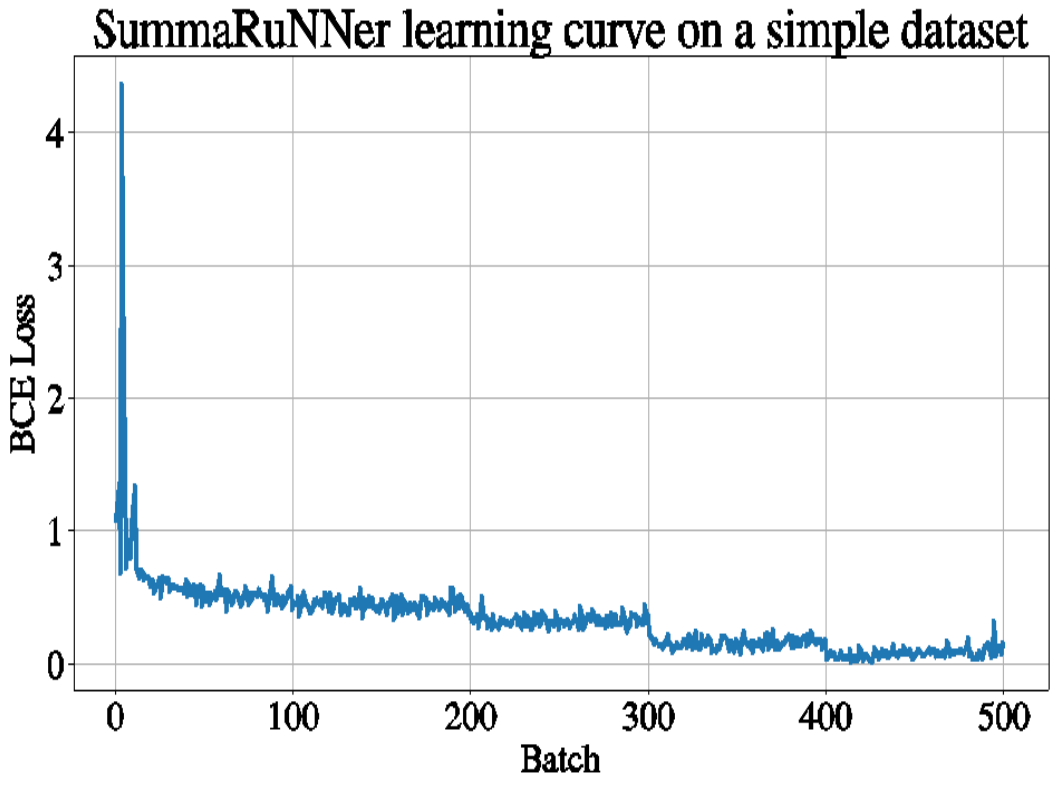
Запуск базового алгоритма
Отчет
Захаров Павел

Согласно заданию требовалось запустить базовые составляющие исследуемой задачи. В качестве таковых использовались модель реферирования текста SummaRuNNer, описанная в [1] и реализованная в [2], а также модель машинного перевода openNMT [3].

SummaRuNNer реализован на языке Python с использованием библиотеки PyTorch. Уровень слов принимает векторы размерности 100 и имеет скрытое состояние размерностью 200. Уровень предложений принимает вход размерности 400 и также имеет скрытое состояние размерности 200. Линейные и билинейные слои, представляющие различные характеристики предложений, также имеют входы длины 400.

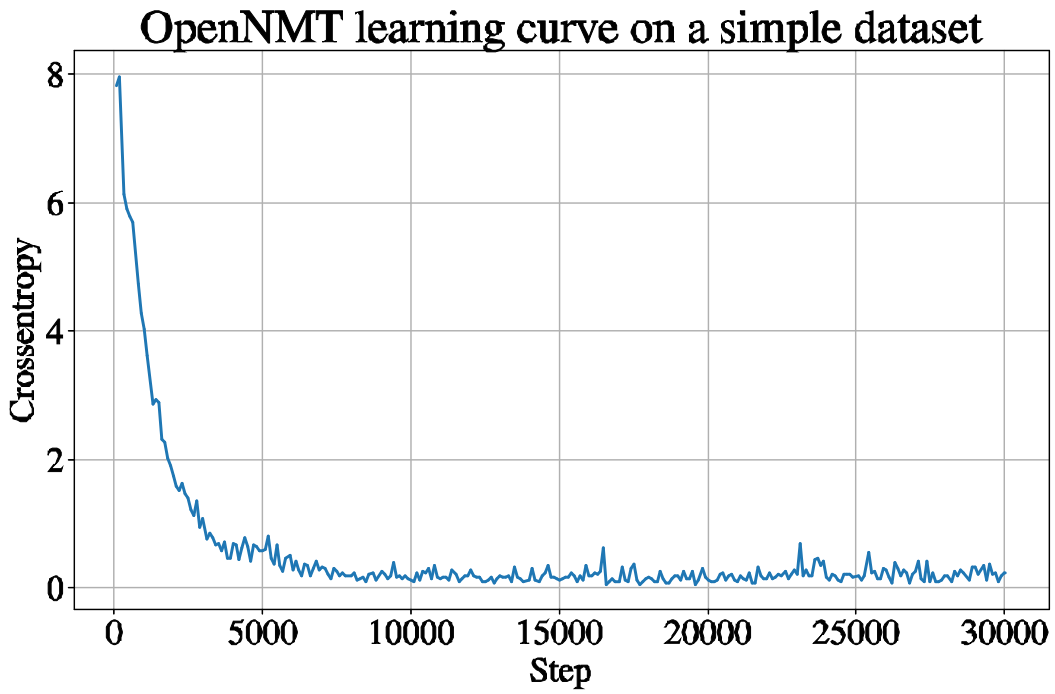
В качестве данных для обучения и валидации используется сокращенный датасет CNN/DailyMail [4] — объем выборки сокращен до 1000 документов. Произведено обучение в 5 итераций, batch_size=10. Кривая обучения:



После этого проведена валидация обученной сети. С использованием полученных оценок предложений построены сокращенные тексты. Между ними и известными сокращенными документами посчитана метрика ROUGE:

Метрика	Recall
ROUGE-1	0,2415
ROUGE-2	0,1031
ROUGE-L	0,1336

Для обучения openNMT использовался сокращенный параллельный подкорпус OPUS: параллельная выборка фраз с онлайн-словаря Tatoeba [5]. Кривая обучения:



Ссылки:

1. R. Nallapati, F. Zhai, and B. Zhou, «Summarunner: A recurrent neural network based sequence model for extractive summarization of documents» CoRR , vol. Abs/1611.04230, 2016.

2. https://github.com/hpzhao/SummaRuNNer

3. https://github.com/OpenNMT/OpenNMT-py

4. https://github.com/deepmind/rc-data

5. http://opus.nlpl.eu/Tatoeba.php