КУРСОВАЯ РАБОТА
Исследовательский проект
“Сжатие словарей для нейросетевого анализа исходных кодов программ”

Выполнил: Андрей Гусев
Руководитель КР: Чиркова Надежда Александровна

Высшая Школа Экономики

aagusev_2@edu.hse.ru

8 июня 2020 г.

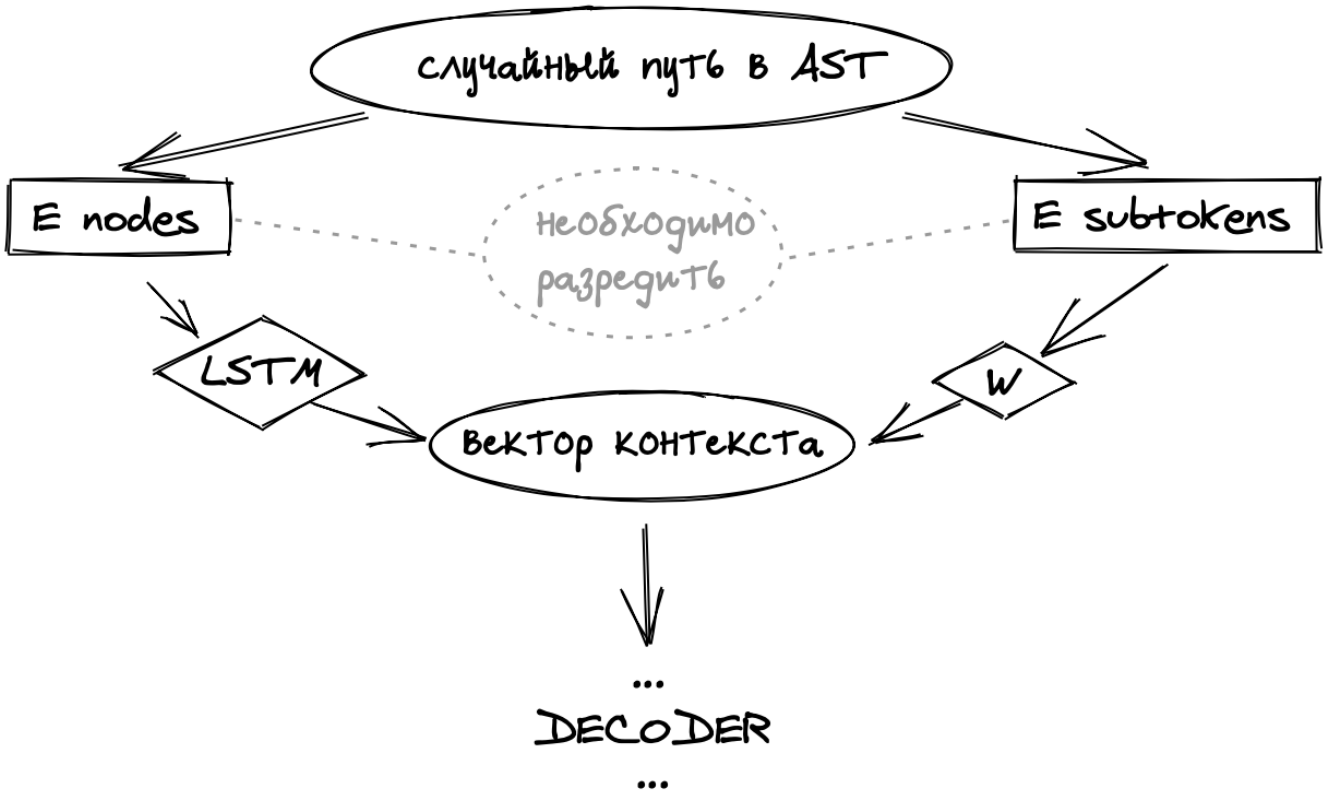
Прунинг входного слоя

Матрица эмбеддингов E задает представление L элементов словаря. Сумму рассматриваемых регуляризаторов $\mathrm{R}(\mathrm{E})$ можно представить в следующем виде:

$$\mathrm{R}(\mathrm{E}) = \underbrace{\lambda_1 \sum_{l=1}^{L} \|\mathrm{E}_l\|_2}_{\text{group Lasso}} + \underbrace{\lambda_2 \|\mathrm{E}\|_1}_{\text{Lasso}}, \text{ где } \|\cdot\|_2 \text{ — это } \ell^2\text{−норма}$$

Lasso-регуляризация стимулирует обнуление одного веса в строке, а group Lasso — к последующему обнулению всей строки.

Encoder Code2Seq



Сравнение с базовой моделью

	Val F1	Test F1	nodes	subtokens
разреженная[1]	0.4239	0.4195	177	10029
ограниченная[2]	0.4201	0.4277	177	10029
оригинальная[3]	0.4092	0.4229	323	73906

[1]Средние результаты трех запусков с предложенной техникой разреживания

[2]Результаты запуска модели с простой техникой сжатия

[3]Усредненные результаты исходной модели