Автоматическая расстановка ударений в словах

Выполнил Захаров Александр Сергеевич

Научный руководитель: Черняк Екатерина Леонидовна
к.т.н., доцент

Постановка задачи

- **Цель работы**: построение модели для расстановки ударений в словах русского языка.

$$\text{слово} \rightarrow \text{сло́во}$$
$$\text{нет руки} \rightarrow \text{нет руки́}$$
$$\text{мои руки} \rightarrow \text{мои ру́ки}$$

- **Актуальность**
  - Синтез речи
  - Транслитерация
  - Изучение русского языка

Существующие подходы к расстановке ударений

- Метод на основе ранжирования [1]
  Accuracy score: 0.839.
- Метод на основе конечного преобразователя [2]
  Accuracy score: 0.962.
- Нейросетевой подход [3]
  Accuracy score: 0.979.
  Эту модель мы будем рассматривать в качестве базовой.

Используемые данные и метрики

- **Данные:**
  - **Источниик:** База данных Акцентологического корпуса Национального корпуса русского языка [4].
  - После обработки 3285455 слов.
  - **Пример:** Давне́нько не бра́л я́ ша́шек в ру́ки! То́ е́сть? А́х не ну́жен жереб\'ец понима́ю тогда́ купи́ у меня́ ка́рую кобы́лу.
- **Метрики:** Accuracy score

$$ACC = \frac{CorrectWords}{AllWords}$$

Общие результаты

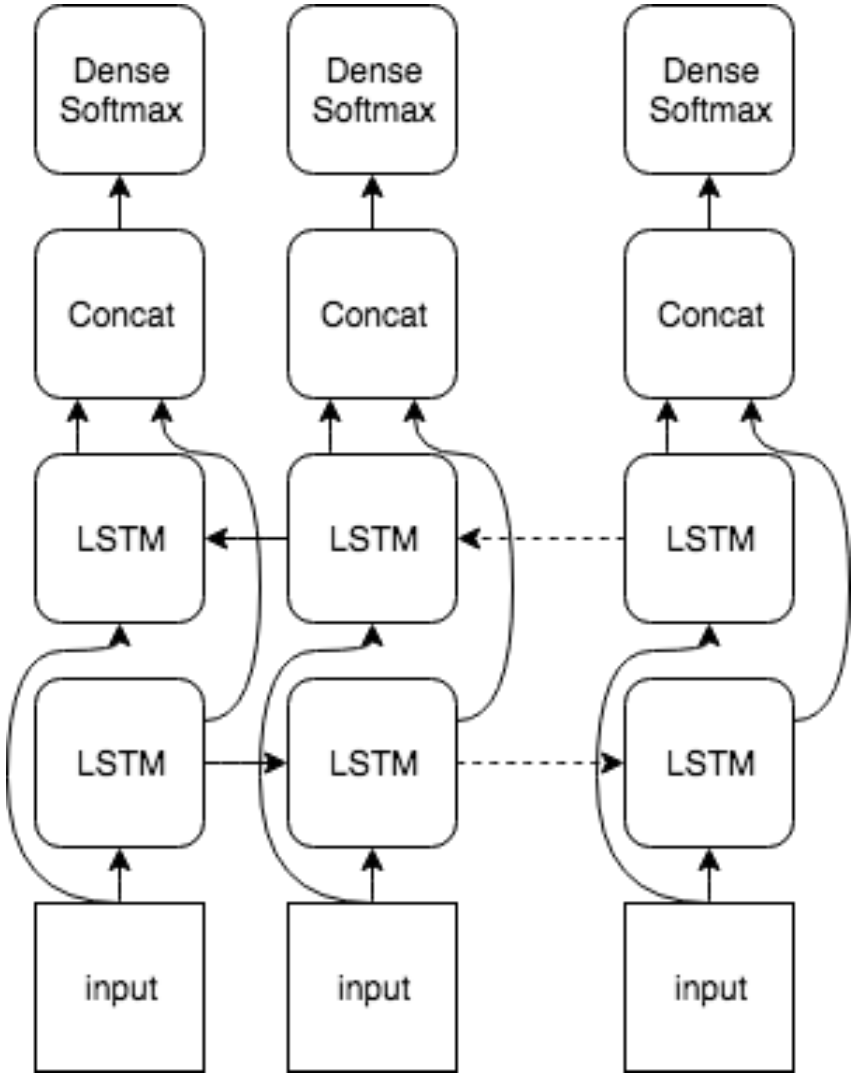
Результаты на тестовой выборке

Модель \ Данные	Символы	Слоги
Локальная		
Глобальная	0.979	
Attention		

Результаты на омографах

Модель \ Данные	Символы	Слоги
Локальная		
Глобальная	0.819	
Attention		

Локальная архитектура



Символьные представление

Пример данных в символьном представлении

Исходная фраза	поднятой руки было
Фраза	ятой руки было
Матрица ответа	11111111011111 00000000100000

Результаты локальной символьной модели

Число слогов	Gl Char	Loc Char
Все слова		
2	0.983	0.961
3	0.977	0.940
4	0.976	0.947
5	0.977	0.960
6	0.973	0.958
7	0.955	0.924
8	0.923	0.866
9	0.952	0.809
AVG	0.979	0.952
Омографы		
2	0.810	0.839
3	0.844	0.774
4	0.847	0.787
AVG	0.819	0.821

Общие результаты

Результаты на тестовой выборке

Данные / Модель	Символы	Слоги
Локальная	0.952	
Глобальная	0.979	
Attention		

Результаты на омографах

Данные / Модель	Символы	Слоги
Локальная	0.821	
Глобальная	0.819	
Attention		

Работа с предложениием

Пример данных для модели по предложениям

Исходная фраза	позволяет добиться того
Фраза	`позволяет добиться того`
Матрица ответа	`111111011111101111111100000001000000100000001`

Результаты локальной модели по предложениям

Число слогов	Loc Char	Loc Sent
Все слова		
2	0.961	0.897
3	0.940	0.891
4	0.947	0.902
5	0.960	0.927
6	0.958	0.925
7	0.924	0.898
8	0.866	0.855
9	0.809	0.647
AVG	0.952	0.898
Омографы		
2	0.839	0.831
3	0.774	0.754
4	0.787	0.775
AVG	0.821	0.812

Слоговое представление

- Ударение падает только на гласные
- В каждом слоге 1 гласная
- **Правила разделения:** Всегда разделяем после гласной буквы, кроме
  - последнего слога (за-бор)
  - пары й + согласная (вой-на)
  - пары непарная согласная (р, л, м, н) и парная согласная (кар-тон)

Пример данных в слоговом представлении

Исходная фраза	позволяет добиться того
Фраза	ля_ет до_би_ться то_го
Матрица ответа	1 1 1 1 0 11 1 1 0 0 0 0 1 00 0 0

Результаты локальной слоговой модели

Число слогов	Gl Char	Loc Char	Loc Syl
Все слова			
2	0.983	0.961	0.985
3	0.977	0.940	0.972
4	0.976	0.947	0.972
5	0.977	0.960	0.976
6	0.973	0.958	0.977
7	0.955	0.924	0.947
8	0.923	0.866	0.899
9	0.952	0.809	0.843
AVG	0.979	0.952	0.978
Омографы			
2	0.810	0.839	0.889
3	0.844	0.774	0.832
4	0.847	0.787	0.843
AVG	0.819	0.821	0.877

Общие результаты

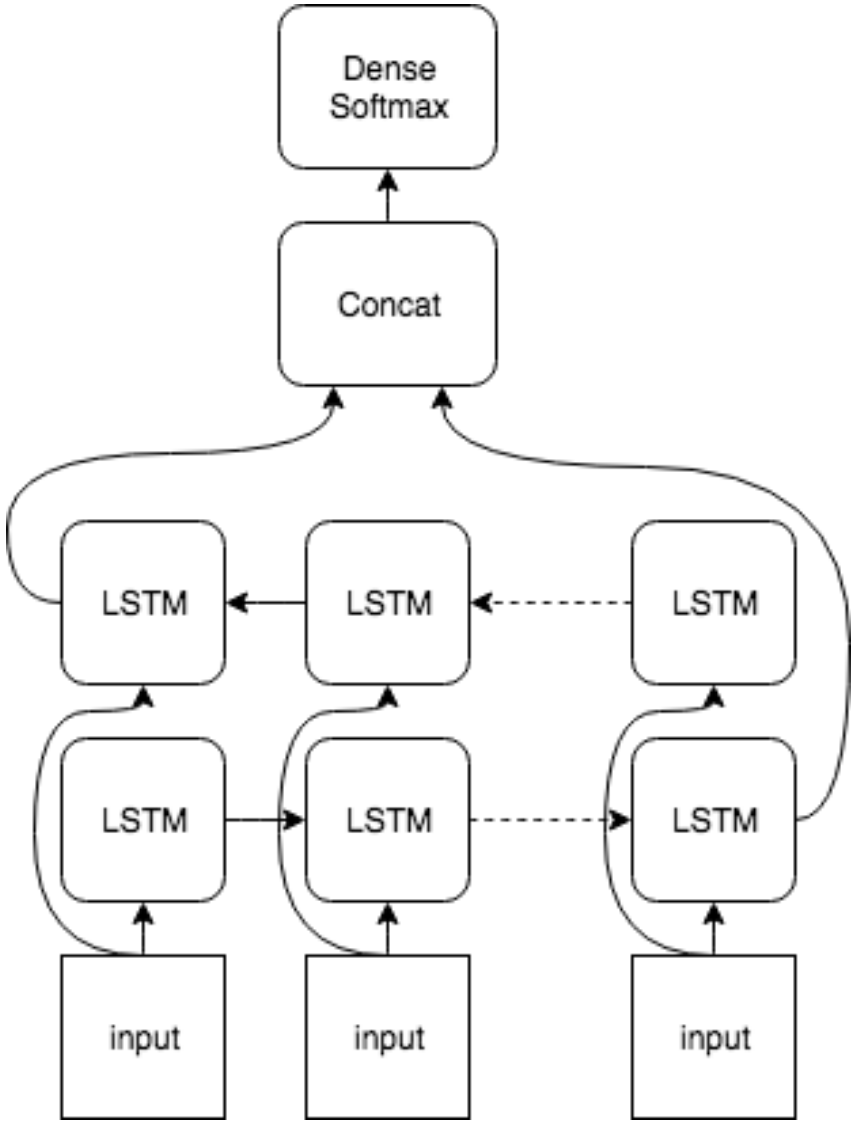
Результаты на тестовой выборке

Данные / Модель	Символы	Слоги
Локальная	0.952	0.978
Глобальная	0.979	
Attention		

Результаты на омографах

Данные / Модель	Символы	Слоги
Локальная	0.821	0.877
Глобальная	0.819	
Attention		

Глобальная архитектура



Результаты глобальной модели

Число слогов	Gl Char	Loc Syl	Gl Syl
Все слова			
2	0.983	0.985	0.985
3	0.977	0.972	0.978
4	0.976	0.972	0.977
5	0.977	0.976	0.977
6	0.973	0.977	0.970
7	0.955	0.947	0.945
8	0.923	0.899	0.895
9	0.952	0.843	0.849
AVG	0.979	0.978	0.981
Омографы			
2	0.810	0.889	0.893
3	0.844	0.832	0.847
4	0.847	0.843	0.852
AVG	0.819	0.877	0.882

Общие результаты

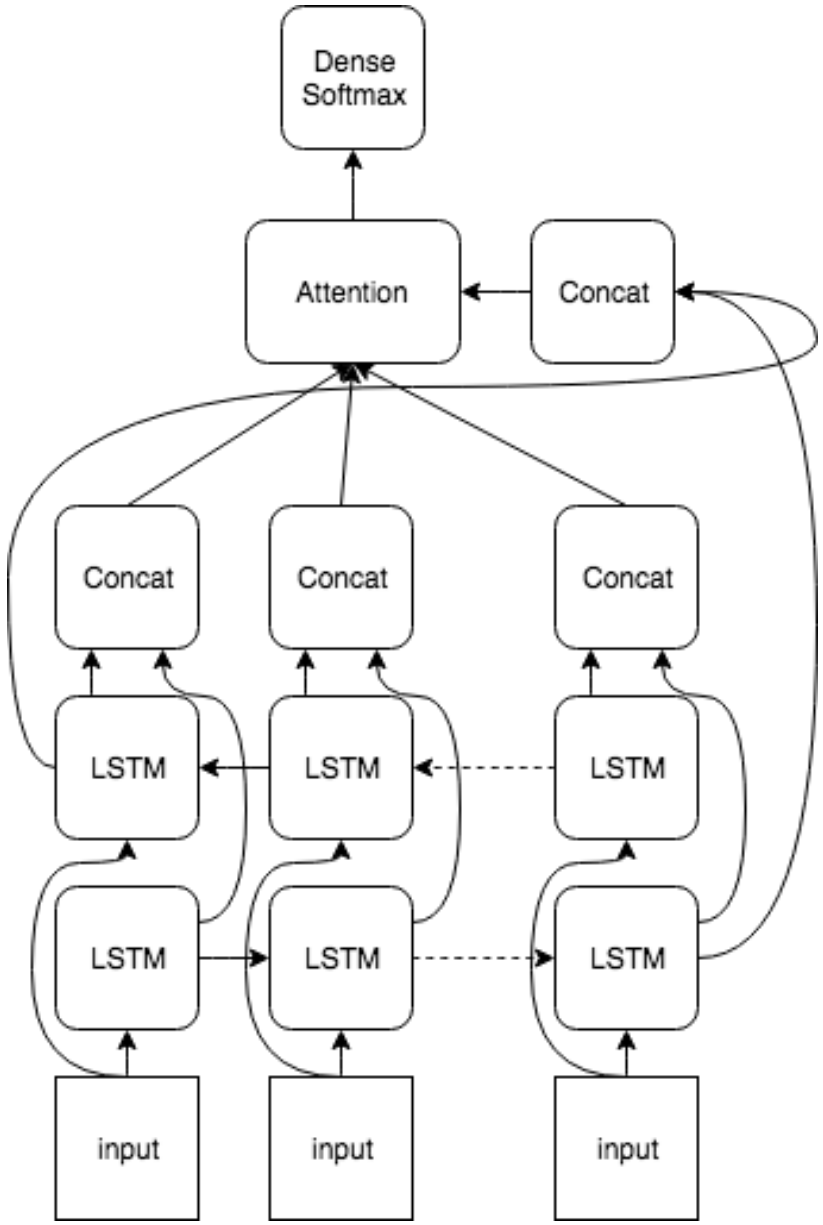
Результаты на тестовой выборке

Данные / Модель	Символы	Слоги
Локальная	0.952	0.978
Глобальная	0.979	0.981
Attention		

Результаты на омографах

Данные / Модель	Символы	Слоги
Локальная	0.821	0.877
Глобальная	0.819	0.882
Attention		

Модель с Attention



Результаты модели с Attention

Число слогов	Gl Syl	Att Syl
Все слова		
2	0.985	0.989
3	0.978	0.982
4	0.977	0.979
5	0.977	0.980
6	0.970	0.969
7	0.945	0.936
8	0.895	0.867
9	0.849	0.747
AVG	0.981	0.985
Все слова		
2	0.893	0.900
3	0.847	0.869
4	0.852	0.846
AVG	0.882	0.889

Общие результаты

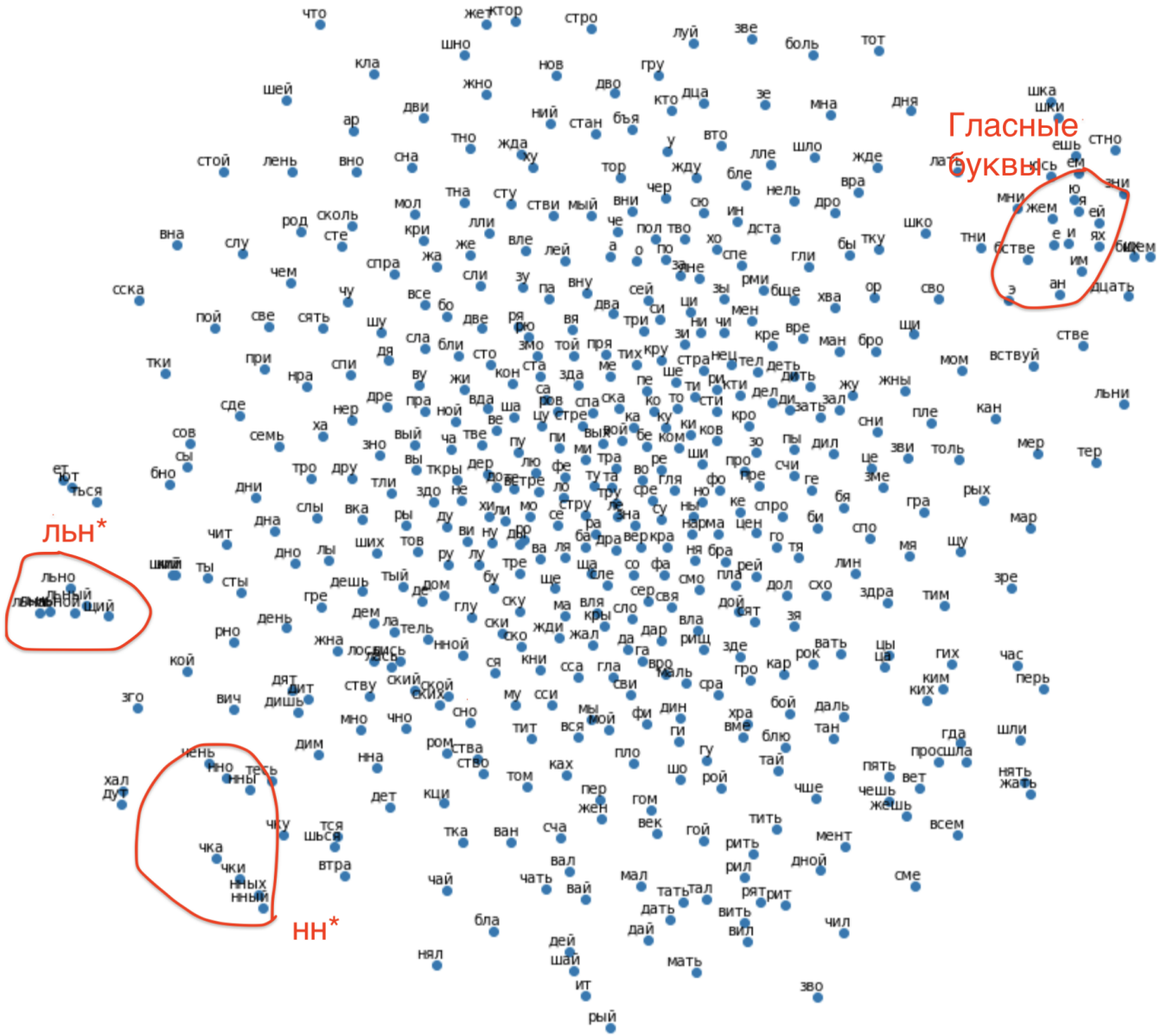
Результаты на тестовой выборке

Данные / Модель	Символы	Слоги
Локальная	0.952	0.978
Глобальная	0.979	0.981
Attention		0.985

Результаты на омографах

Данные / Модель	Символы	Слоги
Локальная	0.821	0.877
Глобальная	0.819	0.882
Attention		0.889

Векторные представлениия слогов



Анализ ошибок: омографы

- 5% в тексте
- 15% среди ошибок
- Можно разделить на группы
  - **Словоформы**
    рукú - рýки
    Accuracy score: 0.85
  - **Разные части речи**
    ужé -ýже
    Accuracy score: 0.94
  - **Одна часть речи**
    лéкартсво - лекáрство
    Accuracy score: 0.76

Анализ ошибок: влияние контекста

Зависимость от наличия контекста

Тип контекста	Accuracy score
Левый и правый	0.986
Левый	0.984
Правый	0.977
Без контекста	0.976

Распределение весов attention

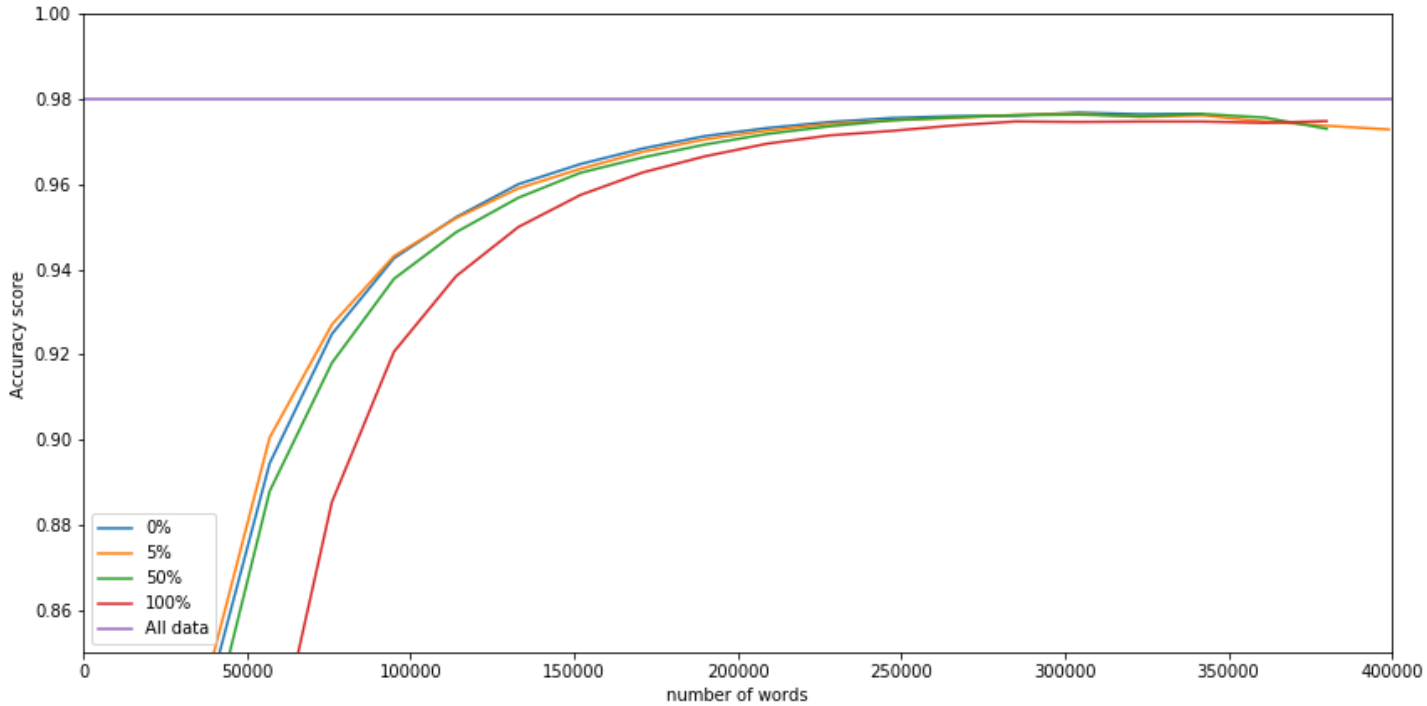
Тип слов	Левый контекст	Левый пробел	Слово
Все слова	0.008	0.294	0.551
Омографы	0.014	0.455	0.391
Тип слов	Правый контекст	Правый пробел	
Все слова	0.005	0.140	
Омографы	0.007	0.130	

Анализ ошибок: новые слова

- 10000 слов
- Accuracy score: 0.838
- Основные ошибки:
  - Русские фамилии
  - Заимствованные слова
  - Многоосновные слова

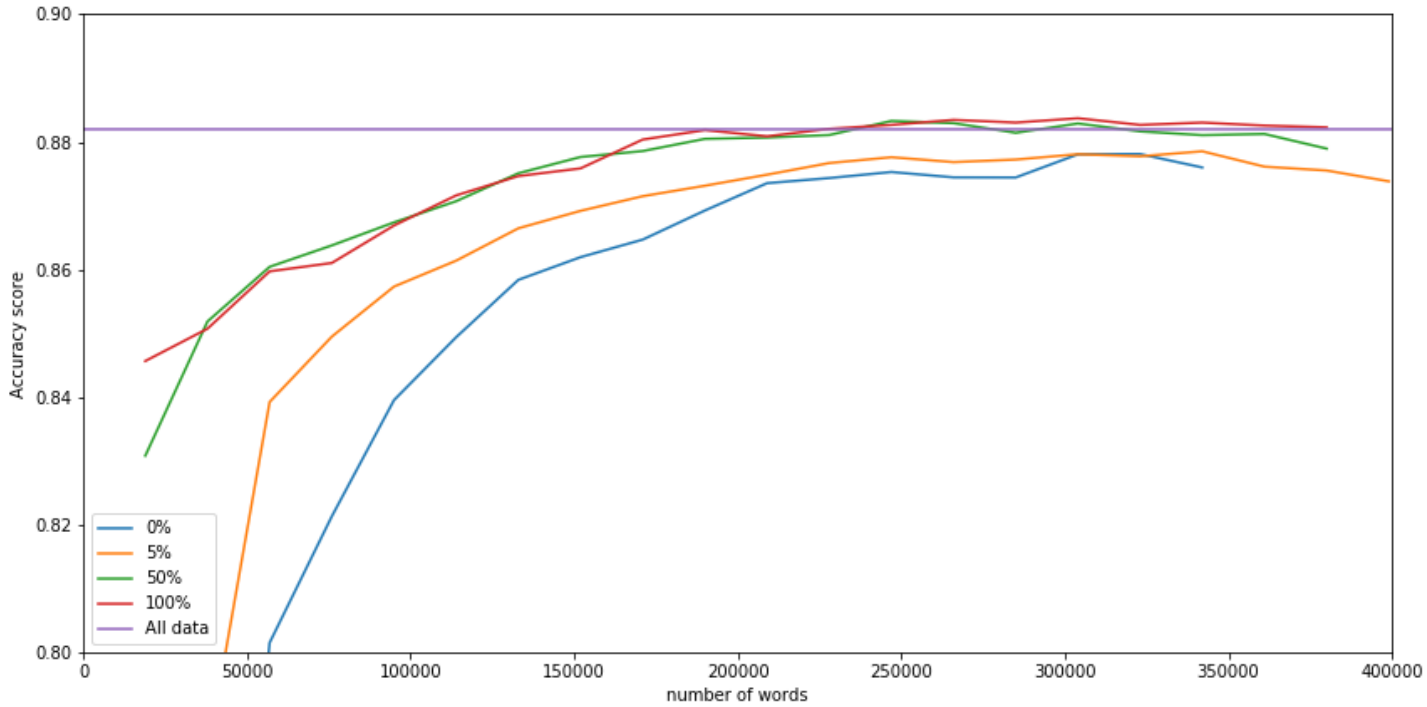
Active learning [5]

Результаты на тестовой выборке



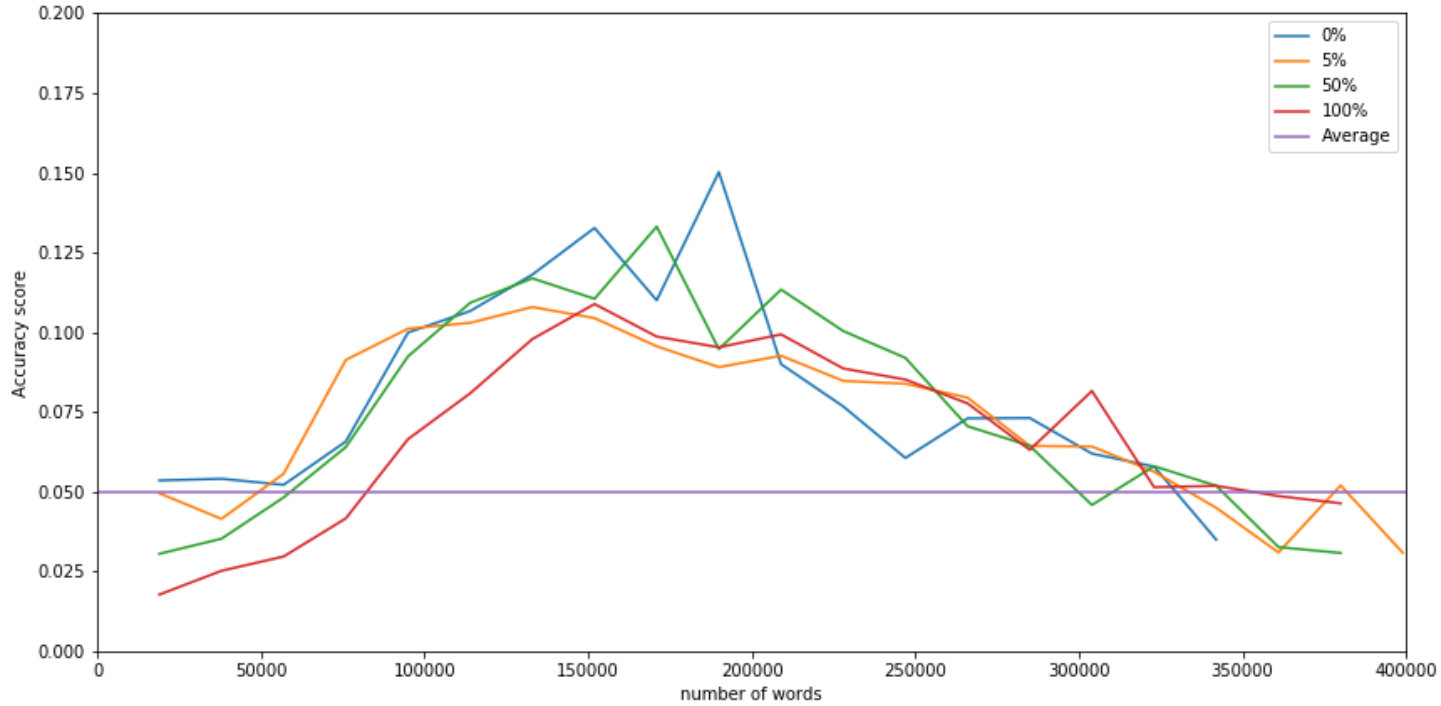
Active learning

Результаты на омографах



Active learning

Доля омографов в новых данных



Заключение

- Модель с attention показала наилучший результат(0.985)
- Слоговое представление данных позволяет улучшить результат с сохранением архитектуры
- При помощи Active learning удалось снизить объем необходимых данных в 5 раз, сохранив при этом качество

Спасибо за внимание

Список литературы

[1] Hall, K. and Sproat, R. (2013). Russian stress prediction using maximum entropy ranking. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 879–883, Seattle, Washington, USA*.

[2] Reynolds, R. and Tyers, F. (2015). Automatic word stress annotation of russian unrestricted text. *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015). Linkoping University Electronic Press, Sweden, Vilnius, Lithuania*.

[3] Ponomareva, M., Milintsevich, K., Chernyak, E., and Starostin, A. (2017). Automated word stress detection in russian. *Proceedings of the First Workshop on Subword and Character Level Models in NLP, 31–35, Copenhagen, Denmark*.

[4] Гришина Е. А. (2009). Корпус «История русского ударения» // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. *СПб.: Нестор-История*.

[5] Shen, Y., Yun, H., Lipton, Z. C., Kronrod, Y., and Anandkumar, A. (2017). Deep active learning for named entity recognition. *Conference paper at ICLR 2018*.

Общие результаты

Результаты на тестовой выборке

Данные / Модель	Символы	Слоги
Локальная	0.952	0.978
Глобальная	0.979	0.981
Attention		0.985

Результаты на омографах

Данные / Модель	Символы	Слоги
Локальная	0.821	0.877
Глобальная	0.819	0.882
Attention		0.889

Сравнение результатов всех моделей

Число слогов	Gl Char	Loc Char	Loc Syl	Gl syl	Att syl
Все слова					
2	0.983	0.961	0.985	0.985	0.989
3	0.977	0.940	0.972	0.978	0.982
4	0.976	0.947	0.972	0.977	0.979
5	0.977	0.960	0.976	0.977	0.980
6	0.973	0.958	0.977	0.970	0.969
7	0.955	0.924	0.947	0.945	0.936
8	0.923	0.866	0.899	0.895	0.867
9	0.952	0.809	0.843	0.849	0.747
AVG	0.979	0.952	0.978	0.981	0.985
Омографы					
2	0.810	0.839	0.889	0.893	0.900
3	0.844	0.774	0.832	0.847	0.869
4	0.847	0.787	0.843	0.852	0.846
AVG	0.819	0.821	0.877	0.882	0.889