Выпускная квалификационная работа

Афракова Алина Владленовна

Методы автоматического морфемного разбора на основе машинного обучения

Научный руководитель:

к.ф.-м.н., доцент Большакова Елена Игоревна

Москва, 2022

Задача морфемного разбора

- Морфемный разбор — это разбиение (сегментация) слова на морфемы (морфы) — минимальные значащие единицы: *пре - крас - н - ый*
- Разбор с классификацией — выделение морфов и распознавание их типов (корень, приставка, суффикс, окончание):

  пре:PREF/крас:ROOT/н:SUFF/ый:END
- Статистические методы разбора недостаточно точны (60-70%), лучшее качество дают методы на основе обучения с учителем (до 88-91%)
- Для высокофлективного русского языка важен метод разбора словоформ, соответствующая нейросетевая CNN-модель* реализована в 2021 г., но учитывает слова длиной не более 20 символов и требует знание части речи

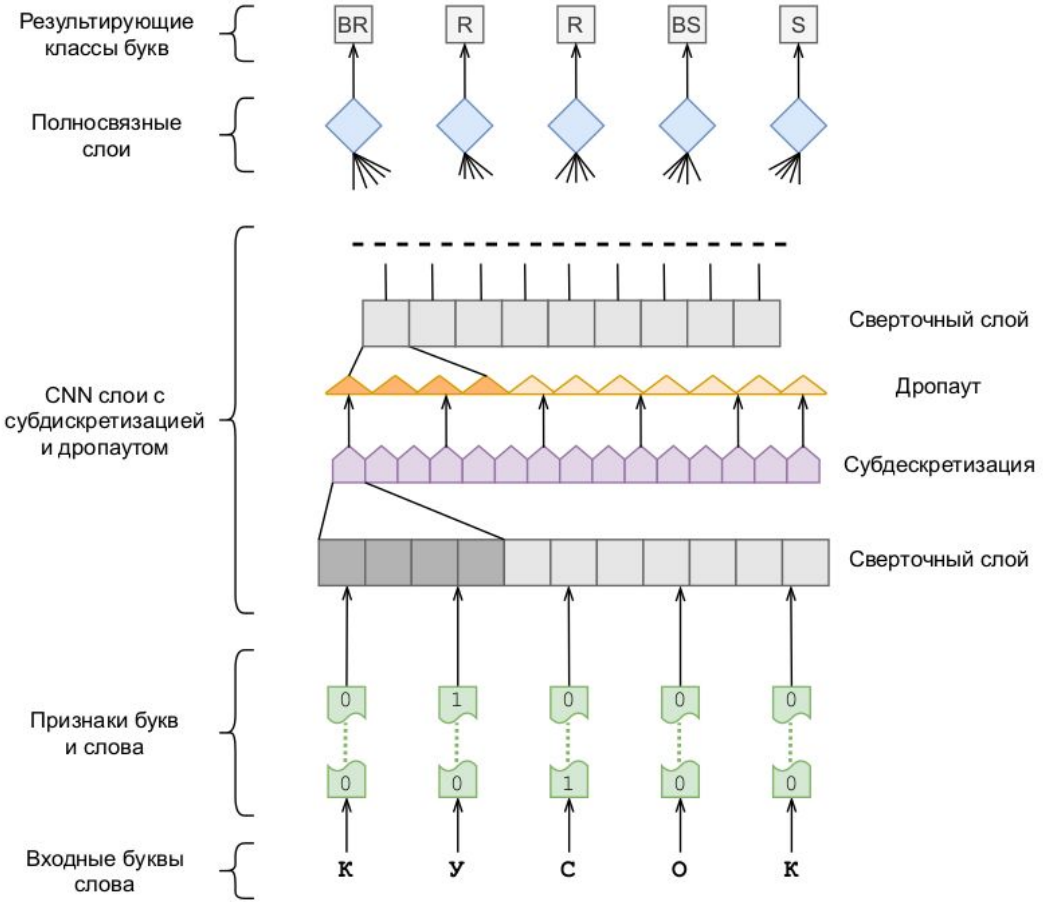
***** *Bolshakova, Sapin, Building Dataset and Morpheme Segmentation Model for Russian Word Forms*

Постановка задачи ВКР

Цель: исследование путей улучшения метода автоматического морфемного разбора словоформ русского языка на базе нейросетевой CNN-модели и наборе размеченных данных RuMorphs-Words

- Изучить подходы к решению задачи морфемного разбора слов и методы на основе машинного обучения, включая указанную CNN-модель
- Разработать дополнительную процедуру морфемного разбора длинных словоформ русского языка
- Программно реализовать нейросетевые модели другой архитектуры за счет добавления новых слоев в данную CNN-модель
- Провести экспериментальные исследования с реализованными моделями морфемного разбора словоформ, оценить их качество работы и производительность (размер моделей и скорость разбора)

Исходная CNN-модель: Архитектура

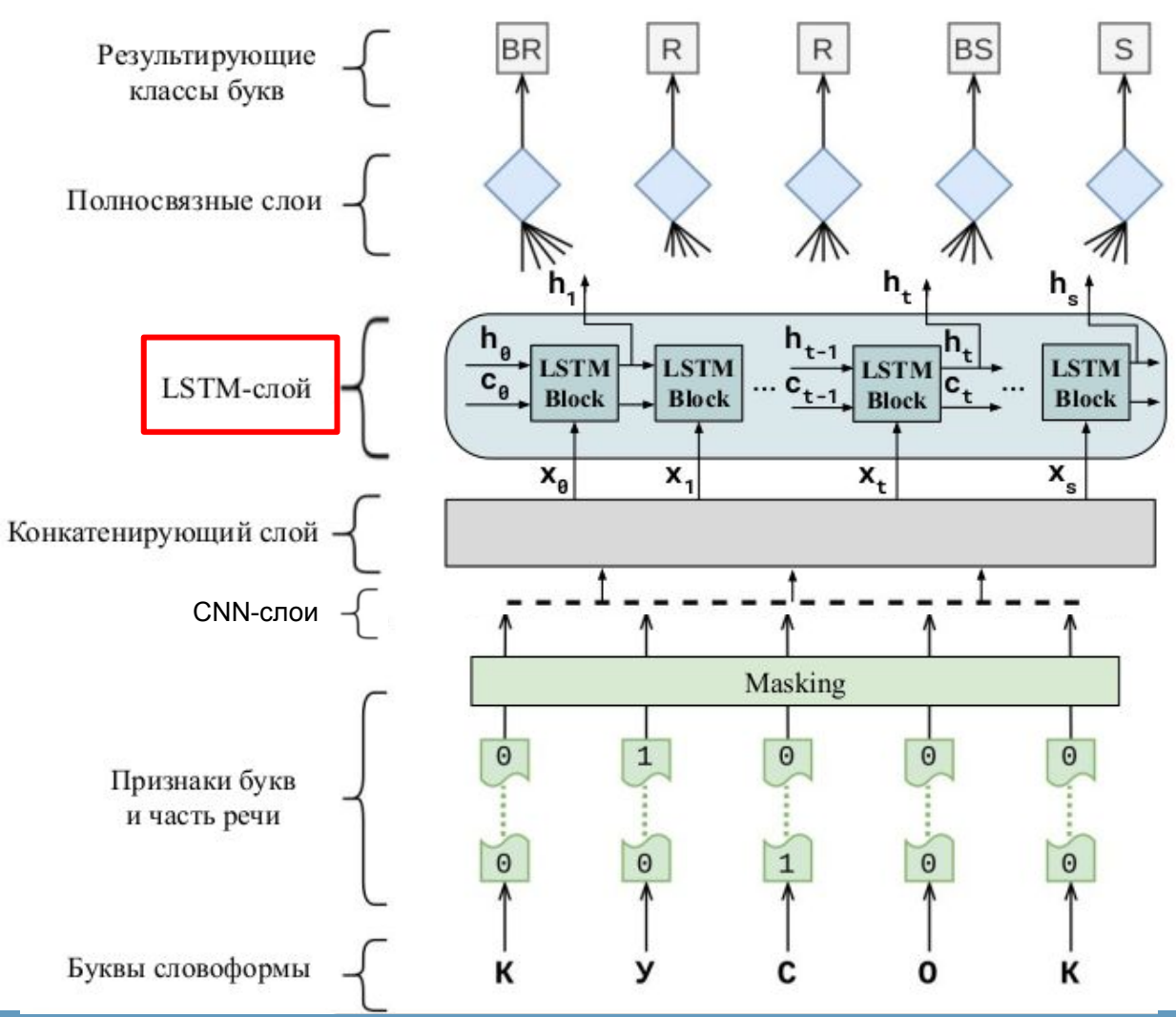


Исходная CNN-модель: Описание

- Задача классификации морфем — классификация / разметка последовательности букв (sequence labeling)
- Входные данные:
  - буквы слова, закодированные в формате one-hot-encoding
  - признак гласности буквы (1 или 0)
  - закодированная часть речи слова
- Выходные данные: результирующий класс буквы
- Целесообразно добавить новые слои: LSTM и Attention
  - Рекуррентные сети, в частности LSTM, хорошо зарекомендовали себя в задачах анализа и обработки временных последовательностей
  - Механизм внимания (Attention) стал неотъемлемой частью в решении задач обработки последовательности и sequence labelling

Морфемный разбор длинных словоформ

- Собрана статистика длинных словоформ (> 20 букв) набора данных RuMorphs-Words: 70.86% — прилагательные через дефис, из них:
  97.88% слов: перед дефисом стоит соединительная гласная:
  *командн**о**-административный*
- Разработана процедура морфемного разбора таких словоформ путем разбиения по дефису и независимой обработки:
  - Часть слова перед дефисом разбирается CNN-моделью как наречие, вторая часть — как прилагательное
  - Результаты морфемного разбора склеиваются и оцениваются
- Полученное значение доли правильных разборов словоформ — до 85%, что сопоставимо с качеством для других словоформ

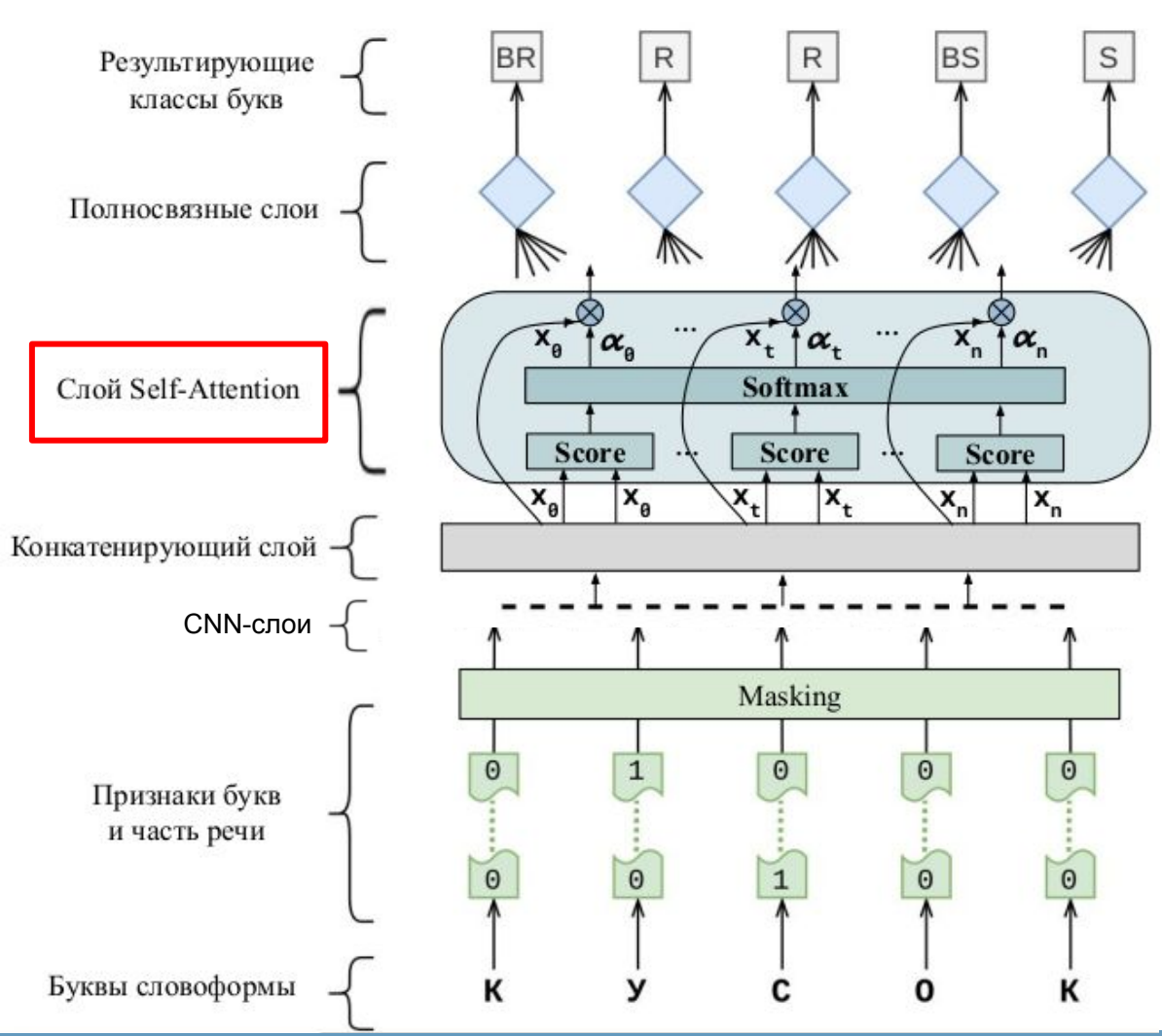


CNN-модель + LSTM: Архитектура

Модель разбора CNN + LSTM

- В исходную нейронную CNN-модель добавлен слой LSTM перед полносвязными слоями
- Модель реализована с использованием языка *Python* и библиотеки *keras*
- В экспериментах определены гиперпараметры для данного слоя и получены оценки качества морфемного разбора
- Accuracy — доля правильных разборов (аккуратность) для букв и слов

Модель	Сегментация			Классификация	
	Точность	Полнота	F-мера	$\text{Accuracy}_{\text{letters}}$	$\text{Accuracy}_{\text{words}}$
CNN	98.21	99.11	98.65	97.72	91.11
CNN + LSTM	98.80	99.08	**98.94**	98.16	**93.00**



CNN-модель + Attention: Архитектура

Модель разбора CNN + Attention

- В CNN-модель добавлен слой Attention перед полносвязными слоями
- Программно реализован слой Attention, язык *Python*, библиотека *keras*
- Экспериментально исследованы варианты моделей с различными размерностями матриц весов: 32, 64 и 128
- Оценено качество разбора и размер моделей

Модели	# параметров сети	Размер модели (МБ)	Точность	Полнота	F-мера	$Accuracy_{words}$
Исходная CNN	2,785,803	**31.93**	98.21	99.11	98.65	91.11
CNN + LSTM	3,622,795	41.52	**98.80**	99.08	**98.94**	**93.00**
CNN + Attention 32	2,884,172	**33.07**	98.60	**99.13**	98.86	**92.55**
CNN + Attention 64	2,982,540	34.19	98.57	99.10	98.83	92.41
CNN + Attention 128	3,179,276	36.45	98.43	**99.15**	98.79	92.20

Реализация слоя Attention

Класс Attention, наследующий базовый класс Layer из модуля *keras.layers* фреймворка *Keras*, основанного на открытой библиотеке *Tensorflow.*

- Класс Attention производит вычисления:

$$score = \tanh(x_t W_t + x_t W_x + b_h) \cdot W_a + b_a$$

$$\alpha = softmax(score)$$

$$\alpha = \alpha \times mask$$

$$output = \alpha \odot x_t$$

где W_t, W_x, W_a — матрицы весовых коэффициентов, b_h, b_a — вектора смещений, *score* — оценочная функция, α — веса внимания, x_t — входной вектор

- Класс Attention включает методы:
  - *build* — инициализация нач. значений матриц весов и смещений
  - *call* — производит основные вычисления по формулам

Сравнение реализованных архитектур моделей

- Ряд экспериментов по оценке:
  - производительности моделей морфемного разбора
  - комбинаций слоев моделей — со слоем конкатенации и без (*no Concat*)
  - моделей, на вход которым уже не подается часть речи слова (*no POS*)
- Итоговая наилучшая модель — *CNN + Attention 32, no POS, no Concat*

Модели	# параметров сети	Размер модели (МБ)	# обрабат. слов в сек.	$\text{Accuracy}_{\text{words}}$
Исходная CNN	**2,785,803**	**31.93**	**377.86**	91.56
CNN + LSTM	3,622,795	41.52	260.25	**93.53**
CNN + Attention 32	2,884,172	33.07	347.90	**92.85**
CNN + Attention 32, no POS	2,830,412	32.45	348.38	92.72
CNN + Attention 32, no Concat	2,807,372	32.19	363.75	**92.98**
CNN + Attention 32, no POS, no Concat	**2,753,612**	**31.57**	**365.12**	92.66

Заключение: результаты работы

- Изучены методы автоматического морфемного разбора слов, нейросетевая CNN-модель для словоформ русского языка, а также рекуррентные нейронные сети, слой LSTM, слой Attention
- Разработана процедура морфемного разбора длинных словоформ русского языка, на основе CNN-модели и анализа набора данных RuMorphs-Words, показавшая достаточно высокое качество
- Программно реализованы высокоточные нейросетевые модели архитектур: CNN + LSTM и CNN + Attention, для автоматического морфемного разбора словоформ русского языка
- Проведены эксперименты с этими нейросетевыми моделями для подбора их гиперпараметров и оценки качества работы, определена модель, оптимальная по качеству разбора и производительности

СПАСИБО ЗА ВНИМАНИЕ!