МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. ЛОМОНОСОВА

На правах рукописи

Сапин Александр Сергеевич

МЕТОДЫ И СРЕДСТВА МОРФОЛОГИЧЕСКОЙ СЕГМЕНТАЦИИ ДЛЯ СИСТЕМ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ

РЕЗЮМЕ
диссертации на соискание учёной степени
кандидата компьютерных наук

Москва — 2023

Диссертационная работа выполнена в федеральном государственном бюджетном образовательном учреждении высшего образования «Московский государственный университет имени М. В. Ломоносова», факультет вычислительной математики и кибернетики.

Научный руководитель: Большакова Елена Игоревна,
кандидат физико-математических наук, доцент,
Московский государственный университет имени М. В. Ломоносова, факультет вычислительной математики и кибернетики, кафедра алгоритмических языков

Общая характеристика работы

Актуальность темы. В настоящее время разработка программных систем для автоматической обработки текстов (АОТ) на естественном языке становится все более востребованным направлением компьютерных наук, причиной этого является активный рост объемов хранимой текстовой информации в электронном виде, в первую очередь, в сети Интернет.

Одними из наиболее используемых программных модулей в составе систем АОТ являются морфологические процессоры, выполняющие морфологический анализ и синтез словоформ текста. Традиционные задачи морфологического анализа включают приведение *словоформы* к нормальной форме (лемме), определение ее морфологических характеристик, а также разрешение морфологической омонимии (неоднозначности характеристик). Например, для словоформы *тетрадей* распознается лемма *тетрадь* и морфологические характеристики: существительное, родительный падеж, множественное число, женский род. Методы решения этих задач морфологического анализа хорошо исследованы, и современные морфопроцессоры решают эти задачи с высоким качеством.

К морфологическому анализу относится также задача *морфологической сегментации*, называемая также *морфемным разбором*, т.е. определение состава слова путем его разбиения (сегментации) на морфы (морфемы), например: *beautiful* → *beauti-ful*, *прекрасный* → *пре-крас-н-ый*. Важность этой задачи определяется тем, что морфемы – минимальные значащие единицы текста, и поэтому могут быть учтены при семантическом анализе текста.

К настоящему моменту известны автоматические методы морфологической сегментации, качество которых недостаточно для приложений АОТ. Решение этой задачи упирается в ряд лингвистических особенностей естественных языков и является особо трудным для языков со сложной морфологией, к каковым относится русский язык (большое количество суффиксов, префиксов, окончаний).

Известны несколько подходов к задаче морфологической сегментации (морфемного разбора) слов. Статистический подход к морфемному разбору был предложен достаточно давно[1], однако он показал достаточно низкую точность решения задачи. В последние годы появились несколько методов на основе машинного обучения[2], повышающих точность решения этой задачи, однако в них рассмотрен лишь частный случай задачи – сегментация нормальных форм слов (лемм). Однако тексты состоят из слов в различной грамматической форме (словоформ), и задача их морфемного разбора требует отдельного изучения. Кроме

[1] Creutz M., Lagus K. Unsupervised models for morpheme segmentation and morphology learning // ACM Transactions on Speech and Language Processing. – 2007 – Vol. 1, no. 1. – C. 1–34.

[2] Ruokolainen T., et al. Painless Semi-Supervised Morphological Segmentation using Conditional Random Fields // Proceedings of the 14th EACL conference. – 2014 – P. 84–89.

того, для применения на практике важен аспект производительности программных реализаций методов морфемного разбора (количество обрабатываемых слов в секунду, потребляемая память), который не исследован вовсе.

К числу задач АОТ, в которых необходима информация о морфемном составе слов, относятся машинный перевод, создание словообразовательных ресурсов (деривационных деревьев и схем порождения новых слов), распознавание смысла новых и редких слов по родственным, а также построение векторных представлений слов (эмбеддингов), которые применяются в подавляющем большинстве современных методов АОТ. Поскольку для словоформ естественного языка высокоточные методы морфологической сегментации еще не созданы, в исследовательских работах используются более простые способы сегментации слов, которые тем не менее повышают качество решения прикладных задач[3]. Для дальнейшего повышения качества необходима более точная информация о внутренней структуре слов, а для этого – разработка и исследование соответствующих методов морфемного разбора.

Поскольку известные морфологические процессоры русского языка не предоставляют возможность морфемного разбора словоформ, также актуально создание процессора, реализующего помимо традиционных функций морфологического анализа текста функцию морфемного разбора слов, на основе высокоточных методов.

Целью данной диссертационной работы является разработка и исследование методов и средств морфологической сегментации слов текста, выполняемой с высокой точностью (качеством) и приемлемой для практики производительностью. Для достижения этой цели необходимо решить следующие **задачи**:

1. Разработать и экспериментально исследовать методы автоматического морфемного разбора нормальных форм слов (лемм) русского языка, реализуемого с высокой точностью (более 88% верно разобранных слов).
2. Разработать метод автоматического морфемного разбора словоформ русского языка с точностью не ниже методов для лемм.
3. Исследовать возможность одновременного решения задачи определения морфологических характеристик и морфемного разбора словоформ русского языка.
4. На основе разработанных методов реализовать модули морфологического процессора, выполняющие функции анализа с достаточной для практики производительностью (более 10 тысяч слов в секунду на одном ядре CPU).

Научная новизна и теоретическая значимость. В данной работе экспериментально исследованы новые методы автоматического морфемного разбора

[3] Hofmann V., Pierrehumbert J., Schütze H. Superbizarre Is Not Superb: Derivational Morphology Improves BERT's Interpretation of Complex Words // Transactions of the Association for Computational Linguistics. – 2021. – Vol. 1. – C. 3594 – 1608.

лемм (нормальных форм) русского языка на основе машинного обучения, среди которых метод на базе сверточной нейронной сети показывает наилучшее качество решения этой задачи (89% верно разобранных слов). Впервые решена задача автоматического морфемного разбора словоформ русского языка, предложенный метод показывает высокое качество разбора словоформ и лемм (90-91%). Также впервые предложен способ одновременного определения морфологических характеристик словоформ и их морфемного разбора, который реализуется с высоким качеством. Данные результаты могут быть применены в качестве базы для построения программных морфологических моделей, распознающих внутреннюю структуру слов, а также могут быть полезны для разработки методов морфологической сегментации текстов на других естественных языках.

Практическая значимость работы состоит в создании программной библиотеки с открытым исходным кодом для морфологического анализа текстов на русском языке, которая:

- Предоставляет функцию морфемного разбора лемм и словоформ русского языка и может использоваться для решения прикладных задач АОТ, в которых одновременно востребованы традиционные функции морфологического анализа, а также морфемного разбора;
- Достигает производительности анализа словоформ до 20 тысяч слов в секунду на одном процессорном ядре для проводимого морфологического анализа, включая морфемный разбор.

Основные положения, выносимые на защиту:

1) Нейросетевой метод автоматического морфемного разбора словоформ русского языка, базирующийся на архитектуре сети, предложенной по результатам исследования методов разбора для нормальных форм слов (лемм). Для реализации метода разработана и применена процедура автоматического построения набора данных с сегментированными словоформами, и показано, что метод для словоформ превосходит по точности известные методы морфемного разбора.
2) Архитектура нейронной сети, на основе которой построен метод одновременного определения морфологических характеристик словоформ текста и их морфемного разбора, с высоким качеством решения обеих задач, а также комплекс моделей морфологического анализа текста, реализующих этот метод и служащих для повышения его производительности.
3) Программная библиотека (морфологический анализатор XMorphy), реализованная с использованием разработанных методов и предназначенная для морфологического анализа и сегментации текстов на русском языке, выполняемых с высокой точностью и производительностью.

Личный вклад. Вышеописанные результаты 2) и 3) получены лично автором диссертации, также он является основным автором работ [1; 2]. Ключевые идеи разработанных моделей морфемного разбора обсуждались и прорабатывались

вместе с научным руководителем Е.И. Большаковой. Также ее вклад состоит в описании правил построения размеченного набора данных с сегментированными словоформами, использованный для обучения модели морфемного разбора словоформ.

Публикации. Основные результаты по теме диссертации изложены в следующих печатных изданиях.

Публикации повышенного уровня

1. Bolshakova E. I., Sapin A. S. A Morphological Processor for Russian with Extended Functionality (*Морфологический процессор для русского языка с расширенными функциями*) // International Conference on Analysis of Images, Social Networks and Texts. – Lecture Notes in Computer Science, V. 10716, Springer, Cham. — 2017. — P. 22—33. — (Scopus, Q2).
2. Bolshakova E. I., Sapin A. S. Building a Combined Morphological Model for Russian Word Forms (*Построение объединенной морфологической модели для словоформ русского языка*) // International Conference on Analysis of Images, Social Networks and Texts. – Lecture Notes in Computer Science, V. 13217, Springer, Cham. — 2022. — P. 45—55. — (Scopus, Q2).

Публикации стандартного уровня

3. Сапин А. С. Построение нейросетевых моделей морфологического и морфемного анализа текста // Труды ИСП РАН. — 2021. — Т. 33, № 4. — С. 117—130. — (список журналов, рекомендованных ВШЭ).
4. Bolshakova E. I., Sapin A. S. Comparing models of morpheme analysis for Russian words based on machine learning (*Сравнение моделей морфемного анализа слов русского языка на основе машинного обучения*) // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference ”Dialogue 2019”. — 2019. — P. 104—113. — (Scopus, без квартиля).
5. Bolshakova E. I., Sapin A. S. Bi-LSTM Model for Morpheme Segmentation of Russian Words (*Bi-LSTM модель для морфемной сегментации слов русского языка*) // Ustalov D., Filchenkov A., Pivovarova L. (eds) Artificial Intelligence and Natural Language. AINL 2019. Communications in Computer and Information Science, V. 1119. Springer, Cham. — 2019. — P. 151—160. — (Scopus, Q3).
6. Bolshakova E. I., Sapin A. S. An Experimental Study of Neural Morpheme Segmentation Models for Russian Word Forms (*Экспериментальное исследование нейросетевых моделей морфемной сегментации для словоформ русского языка*) // Proceedings of the Computational Models in Language and Speech Workshop (CMLS 2020), CEUR Workshop Proceedings. — 2020. — Vol. 2780. — P. 79—89. — (Scopus, без квартиля).
7. Bolshakova E. I., Sapin A. S. Building Dataset and Morpheme Segmentation Model for Russian Word Forms (*Построение набора данных и модели морфемной сегментации для словоформ русского языка*) // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference ”Dialogue 2021”. — 2021. — P. 154—161. — (Scopus, без квартиля).

Апробация работы. Представленные в работе результаты докладывались на следующих международных и российских конференциях, а также семинарах:

1. Научно-технический семинар “Новые информационные технологии в автоматизированных системах”, МИЭМ НИУ ВШЭ, Москва, Россия, 20 апреля 2017 года;
2. Международная конференция “The 6th International Conference on Analysis of Images, Social networks and Texts (AIST 2017)”, Москва, Россия, 27-29 июля 2017 года;
3. Международная конференция “Computational Linguistics and Intellectual Technologies: International Conference Dialogue-2019”, Москва, Россия, 29 мая - 1 июня 2019 года;
4. Международная конференция “Artificial Intelligence and Natural Language. AINL 2019”, Тарту, Эстония, 20-22 ноября 2019 года;
5. Конференция Ломоносовские чтения 2020. Секция вычислительной математики и кибернетики, Факультет вычислительной математики и кибернетики МГУ имени М.В. Ломоносова, Online, 21 октября - 2 ноября 2020 года;
6. Международная конференция “XVI TEL International conference on computational and cognitive linguistics”, Online, 12-13 ноября 2020 года;
7. Международная конференция “Computational Linguistics and Intellectual Technologies: International Conference Dialogue-2021”, Online, 16-19 июня 2021 года;
8. Международная конференция “The 10th International Conference on Analysis of Images, Social Networks and Texts (AIST 2021)”, Тбилиси, Грузия, 16-18 декабря 2021 года;
9. Научный семинар кафедры интеллектуальных информационных технологий факультета вычислительной математики и кибернетики МГУ имени М.В. Ломоносова, 23 декабря 2021 года.

Объем и структура работы.

Диссертация состоит из введения, четырех глав и заключения. Полный объем диссертации составляет 89 страниц включая 20 рисунков и 15 таблиц. Список литературы содержит 84 наименования.

Содержание работы

Во **введении** описывается область исследования, показывается актуальность работы, раскрываются ее цели и задачи, обосновываются научная новизна и практическая значимость работы.

Первая глава посвящена обзору существующих методов морфологического анализа, используемых в современных системах автоматической обработки текста (АОТ).

В **разделе 1.1** введены основные понятия, относящиеся к этапу автоматического морфологического анализа и синтеза текстов на естественном языке, а

также описаны основные метрики, применяемые для оценки качества решения задач морфологического анализа.

В **разделе 1.2** рассмотрены методы морфологического анализа на основе словарей основ и словарей словоформ. Такие методы позволяют решать задачи определения леммы и морфологических характеристик, однако требуют дополнительных средств для разрешения морфологической омонимии, а также отдельных эвристических правил для обработки несловарных слов.

В **разделе 1.3** описываются основные методы разрешения морфологической омонимии (неоднозначности) с использованием словарной и статистической информации и методов машинного обучения – для систем, основанных на словарях.

В **разделе 1.4** рассмотрен подход к морфологическому анализу на основе машинного обучения и векторного представления слов (эмбеддингов), который позволяет полностью избавиться от морфологического словаря и добиться наилучшего качества в задачах лемматизации, определения морфологических характеристик и разрешения морфологической омонимии (до 96.5% правильно определенных лемм и 95% правильно определенных морфохарактеристик для русского языка[4]). Однако методы этого подхода существенно зависят от размеченных данных, на которых производилось обучение, и их реализации обладают низкой производительностью.

В **разделе 1.5** представлен обзор подходов к задаче морфологической сегментации (морфемного разбора) слов естественного языка в двух её вариантах: *морфемная сегментация*, т.е. разбиение слова на составляющие его морфы (например, *зануда* $\rightarrow$ *за* – *нуд* – *а*), и *морфемная сегментация с классификацией*, когда дополнительно к разбиению слова требуется определить типы полученных морфов (например, *зануда* $\rightarrow$ *за*$_{\text{приставка}}$ – *нуд*$_{\text{корень}}$ – *а*$_{\text{окончание}}$).

Для задачи морфемного разбора рассмотрены существующие методы как на основе статистики и обучения без учителя, так и методы на основе обучения с учителем, которые появились лишь в последние годы. Наилучшее качество для русского языка достигается методом на базе сверточных нейронных сетей[5] (88% верно разобранных слов), однако такой метод позволяет обрабатывать только леммы и поэтому не пригоден для обработки текстов (состоящих из словоформ).

В целом, методы морфемного разбора лемм исследованы недостаточно, а для словоформ не разработаны вовсе, тем самым более полное исследование методов морфемного разбора лемм, а также разработка метода, направленного на обработку словоформ, являются актуальными задачами.

---

[4] Lyashevskaya O. N., et al. GRAMEVAL 2020 Shared Task: Russian Full Morphology and Universal Dependencies Parsin // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2020". – 2020

[5] Sorokin A., Kravtsova A. Deep convolutional networks for supervised morpheme segmentation of Russian language // Conference on Artificial Intelligence and Natural Language. – CCIS, Springer, Cham. – 2018

В **разделе 1.6** приведено сравнение предоставляемых функций и производительности (количество обрабатываемых слов в секунду и потребление памяти) свободно доступных морфологических процессоров для русского языка. Показано, что эти процессоры реализуют лишь часть функций морфологического анализа и синтеза, а возможность морфологической сегментации в них отсутствует. Тем самым, актуальна разработка морфопроцессора с поддержкой расширенного набора функций морфологического анализа (включающего морфологическую сегментацию), реализуемых с высоким качеством и производительностью.

Вторая глава посвящена исследованию методов решения задачи морфемного разбора (морфологической сегментации) для лемм русского языка на основе машинного обучения. Из двух вариантов задачи морфемного разбора исследуется более сложная задача сегментации на морфы с классификацией их типов.

В **разделе 2.1** рассмотрены наборы размеченных данных (датасеты) с морфемной разметкой лемм русского языка: RuMorps-Lemmas[6] (96 тысяч лемм) и RuMorphs-CrossLexica (27 тысяч лемм). Описаны особенности датасетов и их разметка, в которой используются 7 типов морфов: PREF (приставка), ROOT (корень), SUFF (суффикс), END (окончание), POSTFIX (постфикс, *ся* и *сь* у глаголов), HYPH (дефис), LINK (соединительная гласная), например, *“бобриха”* – *бобр:ROOT/их:SUFF/а:END*.

В **разделе 2.2** показано, что задача морфемной сегментации с классификацией сводится к задаче классификации последовательности букв сегментируемого слова. В зависимости от используемых классов букв, решается либо задача морфемной сегментации с классификацией групп морфов одного типа (7 классов, при этом последовательные морфы одного типа не отделяются друг от друга), либо задача сегментации с классификацией и разделением последовательно стоящих морфов одного типа (10 классов). Также обосновывается выбор методов машинного обучения для поиска наилучшего решения задачи морфемного разбора лемм: условные случайные поля (CRF), деревья решений с градиентным бустингом (GBDT), рекуррентная нейронная сеть на базе долгосрочно-краткосрочной памяти (LSTM) и одномерная сверточная нейронная сеть (CNN). Результатом обучения на основе конкретного метода является программная модель морфемного разбора.

В конце раздела описаны метрики для оценки качества морфологической сегментации. Для оценки выделения границ морфем применяется точность (*Precision*), полнота (*Recall*) и F1-мера:

$$Precision = \frac{TP}{TP + FP};\ Recall = \frac{TP}{TP + FN};\ F1 = \frac{2TP}{2TP + FP + FN} \quad (1)$$

где TP – количество верно обнаруженных границ между морфами, FP – количество ложно обнаруженных границ, FN – количество не обнаруженных границ.

[6] https://cmc-msu-ai.github.io/NLPDatasets/

Для оценки правильности классификации букв сегментируемых слов используется метрика $Accuracy$:

$$Accuracy_{letters} = \frac{\sum_{i=0}^{len(dataset)} \sum_{j=0}^{len(word_i)} correct(letter_j)}{\sum_{i=0}^{len(dataset)} len(word_i)}, \quad (2)$$

где $len(dataset)$ – количество слов в датасете, $word_i$ – i-ое слово в датасете, $len(word_i)$ – длина i-го слова, $correct(letter_j) = 1$ только когда класс буквы определен верно, и равно 0 иначе. Также оценивается правильность классификации всех сегментированных морфов по всем словам текста (правильность классификации по словам):

$$Accuracy_{words} = \frac{\sum_{i=0}^{len(dataset)} correct(word_i)}{len(dataset)}, \quad (3)$$

где $len(dataset)$ – количество слов в датасете, $word_i$ – i-ое слово в датасете, $len(word_i)$ – длина i-го слова, $correct(word_i) = 1$ только когда определены верно типы и границы всех морфов слова, и равно 0 иначе. Указанная метрика по сути учитывает предыдущие и является основной для оценивания качества морфемной сегментации с классификацией.

Оценка производительности программных моделей морфемного разбора вычисляется как количество обрабатываемых слов в секунду на одном процессорном ядре процессора Intel i7-10850H (на фрагменте коллекции текстов объемом 10 млн слов[7]), также оценивается размер потребляемой памяти в мегабайтах (МБ).

В **разделе 2.3** описывается применение метода условных случайных полей (CRF) для задачи морфемной сегментации с классификацией групп морфов (классификация букв на 7 классов).

В качестве признаков для обучения используются: сама буква, её гласность и некоторые морфологические признаки сегментируемого слова (часть речи, падеж и т.п.), а в качестве данных для обучения и валидации используется датасет RuMorphs-CrossLexica. Результаты экспериментов показывают всего 74.2% правильности классификации по словам.

В **разделе 2.4** описывается применение метода деревьев решений с градиентным бустингом (GBDT) для задачи морфемной сегментации с классификацией букв на 10 классов – для разделения соседних морфов одного типа. Для этого введены 3 класса для начальных букв префиксов, корней, суффиксов, например, для слова *торговец* (*торг*:ROOT/*ов*:SUFF/*ец*:SUFF) получается следующая разметка букв:

Метод GBDT не является методом классификации последовательностей, поэтому используются окна фиксированного размера (5 букв слева и справа от обрабатываемой), остальные признаки обучения совпадают с признаками,

[7] librusec.pro (фрагмент по ссылке https://bit.ly/3typZ57)

т	о	р	г	о	в	е	ц
B-ROOT	M-ROOT	M-ROOT	M-ROOT	B-SUFF	M-SUFF	B-SUFF	M-SUFF

использованными в методе на основе CRF. Для обучения и валидации используются датасеты RuMorphs-Lemmas и RuMorphs-CrossLexica.

Эксперименты с построенной GBDT-моделью показывают высокое качество морфемного разбора для датасета RuMorphs-Lemmas (86.5% правильности классификации по словам) и лучшее качество для RuMorphs-CrossLexica (94.2% правильности классификации по словам). GBDT-модель позволила оценить значимость учитываемых при сегментации признаков: наибольшее влияние на распознавание класса буквы оказывают соседние буквы (предыдущая и две последующих), а среди морфологических характеристик – часть речи.

В **разделе 2.5** приводится описание предложенной архитектуры нейронной сети на основе долгосрочно-краткосрочной памяти (LSTM) для решения задачи морфемной сегментации с классификацией, а также ее обучение и оценка. Поскольку наибольшее влияние на класс буквы (кроме нее самой) оказывают несколько последующих и предыдущих букв, была выбрана двунаправленная LSTM-сеть (BiLSTM). Экспериментальным путем было установлено, что наилучшее качество разбора достигается при использовании многослойной сети (три BiLSTM-слоя) с применением слоев исключения (dropout) между ними и финального полносвязного нейросетевого слоя (рисунок 1).

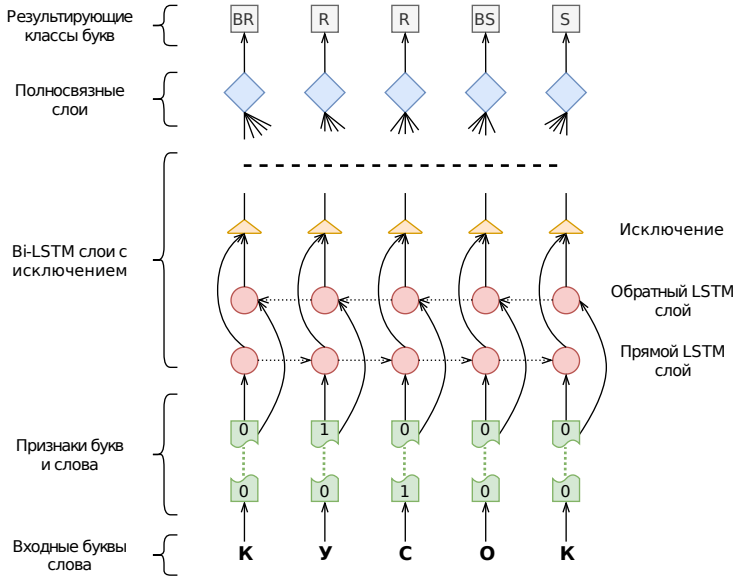

Рис. 1 — Архитектура BiLSTM-модели морфемного разбора

Для обучения сети используются те же признаки букв слова, что и в методе на основе GBDT. Ансамбль из трех аналогичных BiLSTM-моделей показывает качество 89.03% правильности классификации по словам при обучении на датасете RuMorphs-Lemmas и 94.5% при обучении на датасете RuMorphs-CrossLexica.

В **разделе 2.6** описывается архитектура сверточной нейронной сети (CNN) для решения задачи морфемного разбора. В ее основе лежат одномерные сверточные сети с применением исключения между ними и финальным полносвязным слоем для классификации. На вход сети поступают слова из 20 букв, более короткие дополняются пробельными (незначащими) символами, а более длинные делятся на части.

Для обучения сети использовались те же признаки, что и в методе на основе GBDT. Для датасета RuMorphs-Lemmas обученная модель показывает 89.5% правильности классификации по словам, а для датасета RuMorphs-CrossLexica 94.7%. Эти результаты являются наилучшими для морфемного разбора лемм русского языка среди рассмотренных моделей.

В **разделе 2.7** производится сравнение разработанных методов морфемного разбора (программных моделей), с точки зрения правильности классификации по словам, ($Accuracy_{words}$ – таблица 1), и с точки зрения производительности (таблица 2). CNN-модель морфемного разбора показывает наилучшее качество классификации по словам и производительность и превосходит ранее предложенную сверточную модель[8] (показавшую 88.6% правильности классификации по словам).

Таблица 1 — Правильность классификации по словам методов морфемного разбора лемм

Модель	**RuMorphs-Lemmas**	**RuMorphs-CrossLexica**
CRF	-	74.2
GBDT	86.54	94.20
BiLSTM	89.03	94.49
CNN	**89.51**	**94.72**

Описанные в данном разделе методы морфемного разбора предназначены для сегментации лемм (нормальных форм) русского языка, однако при обработке текстов необходимо производить разбор не столько лемм, а различных словоформ. Экспериментальная оценка качества морфемного разбора для словоформ с помощью наилучшей из рассмотренных CNN-модели показала менее 48% правильности классификации слов. Причиной этого, для морфологически сложного русского языка, является существенное различие в морфемной структуре различных словоформ одной леммы, например:

[8] Sorokin A., Kravtsova A. Deep convolutional networks for supervised morpheme segmentation of Russian language // Conference on Artificial Intelligence and Natural Language. – CCIS, Springer, Cham. – 2018

Таблица 2 — Характеристики производительности программных моделей морфемного разбора лемм

Модель	Слов в секунду	Размер модели (МБ)
CRF	47	17
GBDT	269	2651
BiLSTM	64	203
CNN	**673**	**4.7**

разбор леммы: *расшить – рас:PREF/ши:ROOT/ть:END*
разбор словоформы: *разошьют – разо:PREF/шь:ROOT/ют:END*
разбор леммы: *лечь – ле:ROOT/чь:END*
разбор словоформы: *ляжет – ляж:ROOT/ет:END*

Таким образом, для практического применения морфемного разбора необходим метод, предназначеный для сегментации словоформ.

В **третьей главе** описываются разработанные методы морфемного разбора словоформ русского языка, а также созданные для этого наборы размеченных данных (ранее отсутствующих).

В **разделе 3.1** описывается автоматическая процедура построения набора данных (датасета), необходимого для обучения модели морфемного разбора словоформ. На вход процедуре подаются леммы и их морфемные разборы из датасета RuMorphs-Lemmas. Для каждой леммы, с помощью морфологических словарей процедура генерирует словоформы и определяет их часть речи. Для последующей сегментации всех сгенерированных словоформ процедура использует часть речи, а также грамматическую информацию о формообразующих суффиксах и окончаниях русского языка.

Построенный датасет RuMorphs-Words[9] содержит 2.8 миллиона словоформ с морфемной разметкой, в том числе 28% существительных, 45% прилагательных и причастий, 27% глаголов и 0.05% наречий.

В **разделе 3.2** рассматривается метод морфемного разбора словоформ русского языка. Поскольку при сравнении моделей морфемного разбора лемм наилучшее качество и производительность были показаны CNN-моделью, то в качестве основы метода для словоформ была взята аналогичная архитектура нейронной сети. Для обучения нейронной сети используются признаки букв (сама буква и ее гласность) и часть речи сегментируемой словоформы, содержащаяся в построенном датасете.

Архитектура разработанной программной модели (см. рисунок 2) строилась не только с целью достижения выского качества, но и производительности. В ее основе лежат “сверточные блоки”, состоящие из одномерного сверточного слоя, слоя субдискретизации (*max pooling*) и слоя исключения (*dropout*). Слой субдискретизации позволяет значительно ускорить обучение и последующее

[9] https://cmc-msu-ai.github.io/NLPDatasets/

применение модели, а слой исключения помогает бороться с переобучением. Всего в модели используются три последовательно соединенных “сверточных блока”, выход последнего подается на вход полносвязным слоям сети (для каждой буквы слова свой слой сети).

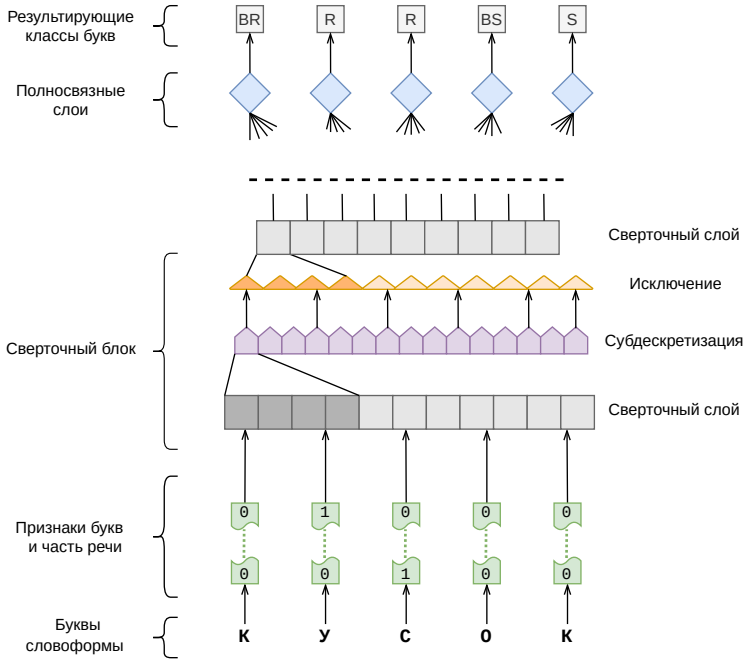

Рис. 2 — Архитектура CNN-модели морфемного разбора словоформ

Обученная на датасете RuMorphs-Words CNN-модель для словоформ показывает 91.06% правильности классификации по словам, при этом на леммах качество также высокое – 90.03% правильности классификации по словам. Модель показала производительность 4559 слов в секунду без учета времени определения части речи сегментируемого слова и 2380 слов в секунду с учетом ее определения. Отдельный шаг определения части речи делает модель не только менее производительной, но и менее удобной в применении, что подводит к задаче разработки метода, выполняющего одновременно морфологический анализ словоформ и их морфемный разбора.

В **разделе 3.3** предлагается архитектура объединенной модели морфологического анализа и морфемного разбора словоформ.

Как и в разработанной CNN-модели морфемного разбора словоформ, в основе архитектуры лежат сверточные нейронные сети, а именно сверточные блоки. В отличие от модели для словоформ, обрабатывающей отдельные слова,

объединённая модель обрабатывает входной текст по предложениям (последовательностям из 9 слов).

Архитектура объединенной модели (рисунок 3) включает подмодель, отвечающую за разрешение морфологической омонимии (слева), а также подмодель, отвечающую за морфемный разбор (справа).

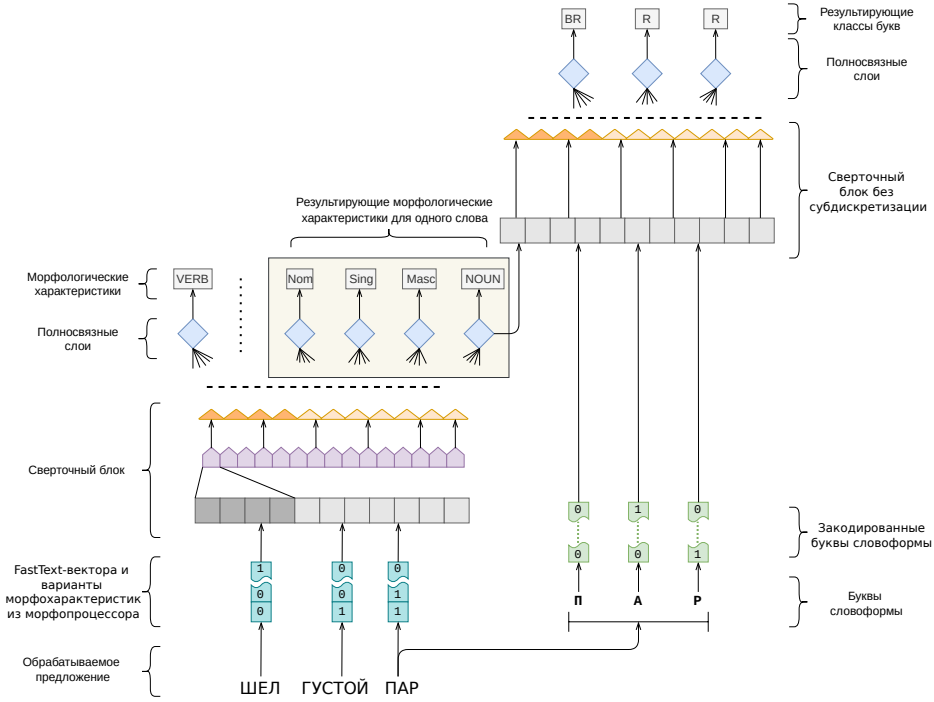

Рис. 3 — Архитектура объединенной морфологической модели

Вместе с каждой словоформой на вход объединенной модели поступают ее возможные морфологические характеристики (которые берутся из морфологического процессора[10]). В случае наличия морфологической омонимии объединенная модель разрешает ее (уточняет часть речи, падеж, число, род, время), а далее уточненная часть речи используется для выполнения морфемного разбора.

Поскольку для обучения объединенной модели необходим размеченный датасет, в котором будет одновременно и морфологическая и морфемная разметка словоформ, а такие датасеты не существовали, то для обучения был взят и дополнительно размечен корпус с морфологической разметкой SynTagRus[11] – в

[10] https://github.com/alesapin/XMorphy

[11] https://universaldependencies.org/treebanks/ru_syntagrus/index.html

нем была добавлена морфемная разметка каждой словоформы. Результирующий датасет получил название RuMorphs-SynTagRus[12].

При обучении объединенной морфологической модели признаками являются возможные варианты морфологических характеристик словоформ, эмбеддинги словоформ, а также закодированные буквы словоформ.

При оценке объединенной модели, обученной на датасете RuMorphs-SynTagRus, было выявлено переобучение (*overfitting*), поэтому далее обучение было разделено на три этапа с применением метода переноса обучения (*transfer learning*). На первом этапе подмодель морфемного разбора обучается на обширном датасете словоформ RuMorps-Words. На втором этапе полученные веса подмодели морфемного разбора замораживаются (исключаются из обучения), и объединенная модель обучается на датасете RuMorphs-SynTagRus. На третьем этапе скорость обучения уменьшается на два порядка и модель целиком обучается на датасете RuMorphs-SynTagRus.

В результате так обученная объединенная морфологическая модель показывает высокое качество разрешения омонимии: 94.2% правильного определения морфологических характеристик. Качество морфемного разбора оказывается наилучшим для датасета RuMorphs-Words (91.7% правильности разбора по словам) и достаточно высоким для датасета RuMorphs-SynTagRus (88.6% правильности разбора по словам).

Производительность модели, реализованной с помощью библиотеки tensorflow-lite, оказалась 1893 слова в секунду, что сравнимо с моделью морфемного разбора словоформ.

В **разделе 3.4** предлагается способ применения (*inference*) разработанных моделей для словоформ. Поскольку нейронные сети (в том числе сверточные) работают с входными данными фиксированного размера, то при обработке текстов зачастую происходит дополнение входных данных (слов и предложений) до фиксированного размера, необходимого для модели: до 20 букв для модели морфемного разбора и до 9 слов для объединенной модели. При этом в реальных текстах на русском языке подавляющее большинство слов содержит меньше 20 букв, и зачастую встречаются предложения короче 9 слов.

Для повышения производительности предлагается использование комплекса из нескольких моделей для различных длин входов: 5, 7, 9, 12 и 15 букв для CNN-модели морфемного разбора, а для объединенной модели – 5, 7, 9 слов и соответственно каждая по 6, 12 и 20 букв. Для каждого очередного входа в зависимости от его размера выбирается наиболее подходящая модель с наименьшим размером.

Применение комплекса моделей значительно ускоряет обработку текста (таблица 3), хотя увеличивает общий размер моделей.

Четвертая глава посвящена описанию библиотечной реализации свободно доступного морфологического процессора русского языка XMorphy[13].

[12] https://cmc-msu-ai.github.io/NLPDatasets/
[13] https://github.com/alesapin/XMorphy

Таблица 3 — Производительность моделей морфемного разбора словоформ

Модель	Слов в секунду	Размер модели (МБ)
CNN	4559	1.1
Объединённая	1893	2.8
CNN (комплекс)	7512	5.4
Объединенная (комплекс)	3543	33.5

В **разделе 4.1** дана общая характеристика процессора, его функции и структура.

XMorphy построен на базе словаря словоформ OpenCorpora[14] с конвертацией морфологических характеристик этого словаря в формат Universal Dependencies[15] (который де-факто становится стандартом для создания размеченных корпусов текстов). Процессор XMorphy поддерживает следующие функции:

– графематический анализ;
– лемматизация и определение морфологических характеристик;
– морфологический синтез;
– разрешение морфологической омонимии;
– морфемный разбор.

Процессор XMorphy реализован в виде библиотеки на языке C++ и набора утилит командной строки (клиент морфопроцессора, утилиты построения словарей). Архитектура исходного кода разделена на логические модули по основным реализуемым функциям (рисунок 4).

Раздел 4.2 посвящен описанию используемых в XMoprhy словарей и структур данных для них. Для хранения более 5 млн. словоформ морфологического словаря используется направленный ациклический граф слов (DAWG[16]), в котором словоформы являются ключами. Для уменьшения размера потребляемой словарем памяти морфологические характеристики сгруппированы в соответствии со словоизменительными классами русского языка, поэтому в качестве хранимых значений используются только номера классов и форм в этих классах, а сами характеристики хранятся в отдельном массиве (рисунок 5).

Морфологический анализ словоформы сводится к её поиску в рассмотренном словаре и возврату всех возможных вариантов морфологических характеристик, а лемматизация выполняется как отсечение окончания формы и приписывания к оставшейся части окончания нормальной формы из словаря. Аналогичным образом реализуется и синтез словоформ.

[14] http://opencorpora.org/
[15] http://universaldependencies.org/
[16] Daciuk J. [et al.]. Incremental construction of minimal acyclic finite-state automata // Computational linguistics. — 2000. — Vol. 26, no. 1. — P. 3—16.

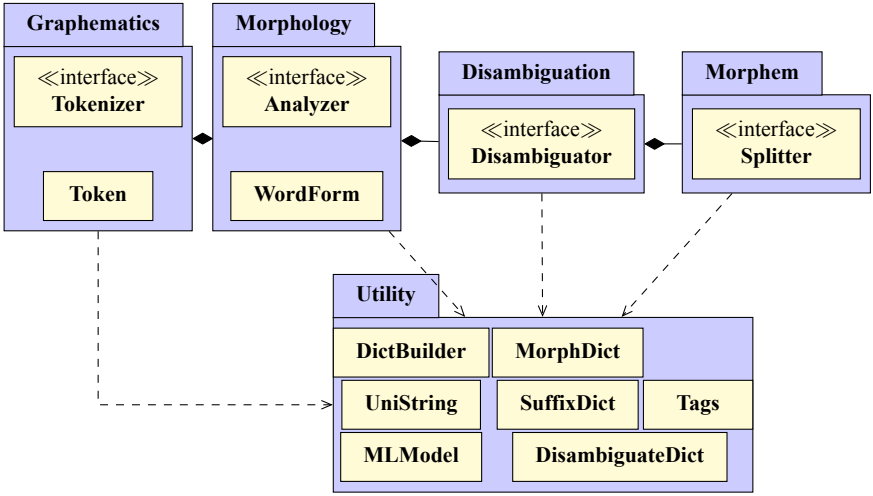

Рис. 4 — Диаграмма модулей процессора XMorphy

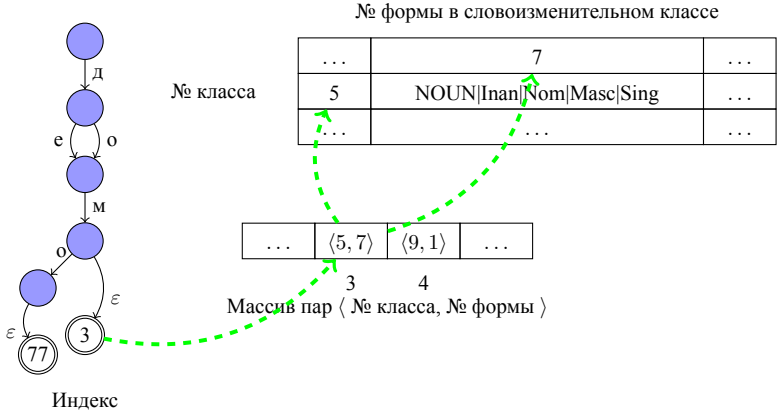

Рис. 5 — Структура морфологического словаря

Для обработки словоформ, отсутствующих в морфологическом словаре, используется отсечение известных приставок и поиск оставшейся части в словаре. Также применяется метод аналогии для определения морфологических характеристик и леммы неизвестной словоформы, например, для неизвестного слова *кринжовая* используется аналогия по окончанию *овая*: *кринж*$\underbrace{\textit{овая}}$ $\rightarrow$ овая (*дворовая, бредовая, ...*) $\underset{\text{ADJ|Nom|Pos|Fem|Sing}}{\rightarrow}$ кринжовая $\rightsquigarrow$ кринжовый $\underset{\text{ADJ|Nom|Pos|Fem|Sing}}{}$

В **разделе 4.3** описываются два реализованных метода разрешения морфологической омонимии. В статистическом (бесконтекстном) методе используется статистика встречаемости каждого набора морфологических характеристик, подсчитанная по размеченному корпусу SynTagRus.

Для снятия омонимии с учетом контекста используется машинное обучение на основе сверточных нейронных сетей. Построенная программная модель аналогичная по архитектуре морфологической подмодели объединенной модели и используется в случаях, когда необходимо выполнить только разрешение морфологической омонимии без морфемного разбора.

В **разделе 4.4** описана реализация модуля морфемного разбора отдельных словоформ и применение объединённой морфологической модели для анализа текстов.

В процессор XMorphy встроен словарь сегментированных словоформ, хранимый в структуре DAWG, содержащий все словоформы датасета RuMorphs-Words: ключом выступает словоформа и часть речи, а значением – её разбор. Если обрабатываемая словоформа не содержится в словаре, то для её морфемного разбора используется разработанный комплекс CNN-моделей морфемного разбора. Значительного ускорения морфемного разбора словоформ удается добиться за счёт использования кеша ранее разобранных слов с алгоритмом вытеснения давно неиспользуемых элементов (LRU).

Комплекс объединенных морфологических моделей применяется в XMorphy для обработки текстов. На вход поступают варианты морфологических характеристик словоформ обрабатываемого предложения (полученные из словаря процессора или предсказанные для несловарных слов) и их FastText-эмбеддинги[17], а также буквы словоформ.

В **разделе 4.5** характеризуются технические особенности реализации XMorphy. Динамическая библиотека, реализующая морфопроцессор, зависит только от стандартных библиотек и может использоваться в любых linux-окружениях. Все словари и модели встраиваются непосредственно в библиотеку, поэтому она может использоваться без предварительной настройки, при этом доступна возможность динамического подключения пользовательских словарей и моделей.

Благодаря использованию архитектуры сверточных нейронных сетей, применению кеширования, эффективной реализации нейронных сетей с помощью библиотеки tensorflow-lite и оптимальных флагов компилятора удается добиться производительности до 20 тысяч слов в секунду для определения морфологических характеристик, леммы и морфемного разбора на одном процессорном ядре процессора Intel i7-10850H.

В заключении приведены основные результаты работы:

1. Предложены четыре метода автоматического морфемного разбора нормальных форм слов (лемм) русского языка на основе различных методов

[17] Bojanowski P., et al. Enriching word vectors with subword information // Transactions of the Association for Computational Linguistics. – 2017. – Vol. 5. – P. 135—146.

машинного обучения, по результатам экспериментальной оценки которых выявлен нейросетевой метод, показывающий наилучшее качество среди всех методов для лемм.
2. Разработана процедура автоматического построения набора данных (датасета) с сегментированными словоформами русского языка.
3. На базе построенного датасета разработан нейросетевой метод для морфемного разбора словоформ, показывающий высокое качество разбора и производительность.
4. Предложена архитектура нейронной сети, на основе которой реализовано одновременное определение морфологических характеристик словоформ текста и их морфемный разбор, с высоким качеством решения обеих задач.
5. С использованием разработанных методов реализована программная библиотека (морфологический процессор XMorphy) для лемматизации, определения морфологических характеристик, разрешения морфологической омонимии, а также морфемного разбора словоформ текстов на русском языке.