MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 32, NO. 4, 2025
journal homepage: www.mais-journal.ru

ARTIFICIAL INTELLIGENCE

Automated Morpheme Segmentation Algorithms for the Belarusian Language: Comparison of Approaches

D. A. Morozov[1], G. O. Feoktistov[1], A. V. Glazkova[2]

DOI: 10.18255/1818-1015-2025-4-384-395

[1]Novosibirsk National Research State University, Novosibirsk, Russia
[2]University of Tyumen, Tyumen, Russia

MSC2020: 68T50
Research article
Full text in Russian

Received September 16, 2025
Revised October 6, 2025
Accepted October 10, 2025

The task of automated morpheme segmentation for morphologically rich but low-resource languages, such as Belarusian, remains insufficiently studied. This paper presents the first large-scale comparative study on the effectiveness of modern neural network approaches to morpheme segmentation using Belarusian language data. We compared three approaches that have demonstrated high quality for other languages: algorithms based on convolutional neural networks (CNNs), algorithms based on LSTM networks, and fine-tuning of BERT-like models. Due to the limited availability of monolingual Belarusian models, we also included larger Russian and multilingual models in the comparison. The experiments were conducted on the openly available Slounik dataset using two strategies for splitting the data into training and test sets. In the first case, the split was random; in the second, words were split by their roots to ensure that words with the same root did not appear in both the training and test sets simultaneously. An ensemble of LSTM networks achieved the best performance in the experiments, with a word accuracy of 91.42 % on the random split and 73.89 % on the root-based split. Comparable results were demonstrated by fine-tuned multilingual and Russian BERT-like models, highlighting the potential of applying large models, including those trained on closely related and higher-resource languages, to this task. An analysis of the errors confirmed that, as with other Slavic languages, the majority of inaccuracies are related to the identification of root boundaries.

Keywords: natural language processing; automated morpheme segmentation; deep learning; Belarusian language; low-resource languages

INFORMATION ABOUT THE AUTHORS

Morozov, Dmitry A. (corresponding author)	ORCID iD: 0000-0003-4464-1355. E-mail: morozowdm@gmail.com PhD, Junior Researcher, Laboratory of Applied Digital Technologies of the International Mathematical Center
Feoktistov, Grigorii O.	ORCID iD: 0009-0001-4486-1270. E-mail: g.feoktistoff@gmail.com Student, Institute for the Humanities
Glazkova, Anna V.	ORCID iD: 0000-0001-8409-6457. E-mail: a.v.glazkova@utmn.ru PhD, Associate Professor, School of Computer Science

For citation: D. A. Morozov, G. O. Feoktistov, and A. V. Glazkova, "Automated morpheme segmentation algorithms for the Belarusian language: comparison of approaches", *Modeling and Analysis of Information Systems*, vol. 32, no. 4, pp. 384–395, 2025. DOI: 10.18255/1818-1015-2025-4-384-395.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 32, № 4, 2025
сайт журнала: www.mais-journal.ru

ARTIFICIAL INTELLIGENCE

Алгоритмы автоматической морфемной сегментации для белорусского языка: сравнение актуальных подходов

Д. А. Морозов[1], Г. О. Феоктистов[1], А. В. Глазкова[2]

DOI: 10.18255/1818-1015-2025-4-384-395

[1]Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[2]Тюменский государственный университет, Тюмень, Россия

УДК 004.912
Научная статья
Полный текст на русском языке

Получена 16 сентября 2025 г.
После доработки 6 октября 2025 г.
Принята к публикации 10 октября 2025 г.

Задача автоматической морфемной сегментации для морфологически богатых, но малоресурсных языков, таких как белорусский, остаётся недостаточно изученной. Настоящая работа представляет собой первое масштабное сравнительное исследование эффективности современных нейросетевых подходов к морфемной сегментации на материале белорусского языка. Мы сопоставили три подхода, показавших высокое качество в случае других языков: алгоритмы на базе свёрточных нейронных сетей, алгоритмы на основе LSTM-сетей и дообучение BERT-подобных моделей. Из-за малого числа доступных моноязычных белорусских моделей, мы также добавили к сравнению более крупные русскоязычные и многоязычные модели. Эксперименты проводились на свободно доступном наборе данных Slounik с использованием двух стратегий разбиения данных на обучающую и тестовую выборки. В первом случае разбиение было случайным, во втором случае слова были разбиты по корням так, чтобы однокоренные слова не могли попасть одновременно в обучающую и тестовую выборки. Наилучшей производительности в ходе экспериментов достиг ансамбль LSTM-сетей с долей полностью верных разборов 91.42 % при случайном разбиении и 73.89 % при разбиении по корням. Сопоставимые результаты продемонстрировали дообученные многоязычные и русскоязычные BERT-подобные модели, что подчёркивает возможность применения в этой задаче крупных моделей, в том числе, обученных на близкородственных и более ресурсообеспеченных языках. Анализ ошибок подтвердил, что большинство неточностей, как и для других славянских языков, связано с определением границ корня.

Ключевые слова: обработка естественного языка; автоматическая морфемная сегментация; глубокое обучение; белорусский язык; малоресурсные языки

ИНФОРМАЦИЯ ОБ АВТОРАХ

Морозов, Дмитрий Алексеевич (автор для корреспонденции)	ORCID iD: 0000-0003-4464-1355. E-mail: morozowdm@gmail.com Канд. тех. наук, младший научный сотрудник Лаборатории прикладных цифровых технологий Международного Математического Центра
Феоктистов, Григорий Олегович	ORCID iD: 0009-0001-4486-1270. E-mail: g.feoktistoff@gmail.com Студент Гуманитарного института
Глазкова, Анна Валерьевна	ORCID iD: 0000-0001-8409-6457. E-mail: a.v.glazkova@utmn.ru Канд. тех. наук, доцент кафедры программного обеспечения Школы компьютерных наук

Для цитирования: D. A. Morozov, G. O. Feoktistov, and A. V. Glazkova, “Automated morpheme segmentation algorithms for the Belarusian language: comparison of approaches”, *Modeling and Analysis of Information Systems*, vol. 32, no. 4, pp. 384–395, 2025. DOI: 10.18255/1818-1015-2025-4-384-395.

Введение

Морфемная сегментация — это разбиение слова на минимальные (неделимые) значимые единицы речи, называемыми морфемами: приставки, корни, суффиксы и другие. Разбиение слова на морфемы необходимо при изучении морфологически развитого языка. Например, в русском корректная морфемная сегментация необходима для правописания -н-/-нн- на границах морфем в слове. В белорусском языке передача звуков [д] и [т] перед мягким звуком [в'] на письме зависит от расположения звуков — если [д] находится на конце префикса, а [т] в составе суффикса, то звуки передаются буквами «д» и «т», иначе аффрикатами «дз» и «ц» соответственно[1]. Наряду с этим, морфемная сегментация применяется в компьютерной лингвистике, например, при токенизации текстов. Недавние исследования показали, что использование морфемно-ориентированной сегментации вместо общепринятых алгоритмов, таких как byte-pair encoding [1], может повысить эффективность языковой модели, в особенности при обработке малоресурсных языков [2—5].

Традиционным источником информации о сегментации слов являются словари морфемных разборов. В то же время, подобные словари значительно уступают реальному количеству слов в языке. Например, в русском языке один из самых больших словарей морфемных разборов «Большой словообразовательный словарь» А. Н. Тихонова насчитывает около 150 тысяч лемм, в то время как Основной корпус Национального корпуса русского языка содержит более 250 тысяч [6]. Принимая во внимание также, что ни один словарь не является исчерпывающим на момент его выхода, автоматическая морфемная сегментация слов становится необходимостью для пополнения словарей.

Задача автоматической морфемной сегментации на сегодняшний день изучена достаточно подробно, а предложенные алгоритмы весьма разнообразны. Среди наиболее актуальных подходов следует упомянуть применение нейросетевых моделей, обучаемых с нуля [7], дообучение предобученных BERT-подобных моделей [8], генерацию сегментаций при помощи больших языковых моделей (LLM) [9, 10]. При этом эффективность подходов меняется в зависимости от конкретного языка и объёма доступной обучающей выборки.

Целью настоящей работы является анализ эффективности подходов к морфемной сегментации на материале белорусского языка. Актуальность работы обеспечивают, с одной стороны, слабая изученность вопроса для белорусского языка, с другой стороны, возможность опираться в выборе методов на результаты для близкородственного и намного более изученного в контексте данной задачи русского языка. В ходе исследования мы сопоставили подходы на базе свёрточных нейронных сетей (CNN), сетей LSTM и дообучения BERT-подобных моделей, предобученных на данных трёх типов: 1) белорусскоязычных, 2) русскоязычных и, наконец, 3) многоязычных. Исследование проводилось на материале открыто доступного аннотированного набора данных Slounik[2], содержащего морфемные разборы для лемм белорусского языка.

В ходе работы были получены следующие ключевые результаты:

- Лучших результатов удалось добиться с использованием ансамбля сетей LSTM с обрезкой градиента и дообучением многоязычной модели google-bert/bert-base-multilingual-cased[3]: доля полностью верных разборов составила 91.42 ± 0.81 % и 91.37 ± 0.39 %, соответственно.
- Как и в случае предыдущих исследований для белорусского и других языков, показано значительное падение качества сегментации при работе со словами, содержащими корни, не встретившиеся в обучающей выборке (out-of-vocabulary, OOV). Доля полностью верных разборов в таком случае оказывается ниже на 17–20 % в сравнении с тестированием на случайной выборке. Лучший результат при этом, как и в случае случайного разбиения, показало

[1] Аб Правілах беларускай арфаграфіі і пунктуацыі. Закон Рэспублікі Беларусь. 23 ліпеня 2008 г. № 420-З

[2] https://huggingface.co/datasets/ruscorpora/morphodict-bel

[3] https://huggingface.co/google-bert/bert-base-multilingual-cased

использование ансамбля сетей LSTM с обрезкой градиента: доля полностью верных разборов составила 73.89 ± 1.06 %.

- В большинстве случаев ошибки автоматической сегментации связаны с определением границ корня, как и в случае других языков. Среди частотных вариантов ошибок встречается сдвиг границы в паре морфем, а также склеивание корня с соседней морфемой или наоборот, избыточное дробление корня.

Настоящая статья состоит из следующих разделов: раздел 1 представляет собой краткий обзор существующих алгоритмов морфемной сегментации; раздел 2 содержит описание использованного датасета, исследованных алгоритмов морфемной сегментации, а также метрик и особенностей тестирования; наконец, раздел 3 содержит результаты проведённых экспериментов, сравнение их с ранее полученными и анализ ошибок подходов, показавших наилучшие результаты.

1. Обзор предметной области

Алгоритмы автоматической морфемной сегментации представлены широким спектром подходов, от правиловых и статистических до нейросетевых. При этом лучшего качества, как правило, удаётся добиться с использованием именно нейросетевых подходов. Это было наглядно продемонстрировано на соревновании SIGMORPHON 2022 Shared Task on Morpheme Segmentation [11], где представленные участниками алгоритмы значительно превзошли правиловые и статистические базовые модели. Наиболее эффективными оказались подходы команд DeepSPIN (модели на базе LSTM-сетей и Transformer) [12] и CLUZH (модели на базе нейронных трансдьюсеров) [13]. Следует отметить крайне высокое качество сегментации, достигнутое в ходе соревнования. Так, F-мера для определения морфем в ходе масштабного тестирования на материале девяти языков, представляющих различные языковые семьи и ветви, у лучших моделей составила от 93 % до 99 % в зависимости от языка. Высокое качество сегментации при помощи нейросетевых подходов, таких как свёрточные нейронные сети и LSTM-сети, подтвердило недавнее исследование на материале семи языков [7]. Исследованы также подходы с использованием предобученных языковых моделей, в том числе, BERT-подобных [8, 14] и LLM [9, 10]. В случае последних авторами получены смешанные результаты: хотя основанные на LLM подходы могут быть сравнительно эффективными для малоресурсных языков, их вычислительная сложность довольно высока [9]. При этом Anderson et al. показали, что LLM-подход может уступать гораздо более простым алгоритмам [10].

Множество исследований в этой области посвящено русскому языку [8, 11—13, 15—19]. В частности, эксперименты проводились с алгоритмами на базе градиентного бустинга над решающими деревьями [16], свёрточными нейронными сетями [15], двунаправленными многослойными LSTM-сетями [17], Transformer-сетями [12], BERT-подобными моделями [8]. Дополнительно исследовалась обобщающая способность алгоритмов при уменьшении обучающей выборки [18] и сегментации слов, содержащих OOV-корни [8, 19]. Было показано, что на случайных словах эффективность автоматической сегментации сопоставима с качеством экспертной разметки [19]. При этом лучшие результаты для русского языка демонстрирует подход с дообучением BERT-подобных моделей, доля полностью верных разборов для которого превышает 92 %.

При этом близкородственный русскому белорусский язык с точки зрения морфемной сегментации исследован крайне слабо. Насколько нам известно, единственной работой, в которой бы исследовалось качество алгоритмов морфемной сегментации на материале белорусского языка, является исследование Morozov et al. [8], в котором на материале трёх славянских языков, в том числе белорусского, тестируются два подхода к морфемной сегментации: с использованием ансамбля свёрточных нейронных сетей и с дообучением BERT-подобных моделей. Достигнутое для белорусского качество сегментации оказалось весьма высоким (доля полностью верных разборов превысила 90 %), при этом, в отличие от чешского и русского, на материале белорусского лучшим подходом оказался именно ансамбль свёрточных сетей, что авторы связали с недостаточно качественными предобу-

Table 1. The characteristics of the dataset

Таблица 1. Характеристики использованного набора данных

Характеристика	Значение
Количество уникальных слов	35 219
Количество уникальных морфем	7 281
Количество уникальных корней	6 530
Средняя длина слова, символов	9.17
Средняя длина слова, морфем	3.77
Средняя длина корня, символов	4.12
Среднее число слов, содержащих некоторую морфему M	18.22
Среднее число слов, содержащих некоторый корень R	5.60

ченными моделями для белорусского языка. Тем не менее, вопрос об эффективности алгоритмов морфемной сегментации на материале белорусского языка нельзя считать достаточно изученным в связи с тем, что данное исследование было ограничено небольшим набором алгоритмов.

2. Данные, модели и методология

2.1. Данные

Источником данных для обучения и тестирования стал массив морфемных разборов Slounik[4], основанный на словаре «Школьны марфемны слоўнік беларускай мовы» [20]. Этот датасет содержит около 35 тысяч морфемных разборов лемм белорусского языка. Каждый разбор представляет собой набор подстрок-морфем, где каждой морфеме приписан один из пяти типов: PREF (приставка), ROOT (корень), SUFF (суффикс), END (окончание), LINK (соединительная гласная). Ранее этот набор данных использовался в исследовании [8], однако при сопоставлении результатов стоит учитывать, что в данном исследовании мы использовали расширенную примерно на 4 тысячи слов версию датасета, что может влиять на эффективность алгоритмов. Основные характеристики датасета кратко перечислены в таблице 1.

2.2. Модели

Все рассмотренные нами модели решают задачу побуквенной классификации: каждому символу в слове присваивается метка, отвечающая за положение символа внутри морфемы и за тип морфемы. При этом рассматривается четыре возможных положения символа внутри морфемы: для морфем, состоящих хотя бы из двух символов — в начале (**B**), в середине (**M**) или в конце морфемы (**E**); для морфем, состоящих из одного символа — единственный символ в морфеме (**S**). В таком случае для слова *рэжысёрскі (режиссёрский)* с морфемным разбором `рэжыс:ROOT/ёр:SUFF/ск:SUFF/і:END` эталонным набором меток будет `[B-ROOT, M-ROOT, M-ROOT, M-ROOT, E-ROOT, B-SUFF, E-SUFF, B-SUFF, E-SUFF, S-END]`.

В соответствии с результатами предыдущих исследований на материале белорусского, русского и других языков [7, 8, 19] мы выбрали для экспериментов три типа алгоритмов:

1. **Алгоритмы на основе свёрточных нейронных сетей.** Мы рассмотрели два варианта алгоритма: трёхслойную свёрточную сеть со 192 фильтрами и размером окна 5, а также ансамбль из трёх таких сетей. В обоих случаях модель обучалась в течение 25 эпох (в случае ансамбля сети обучались последовательно). Мы использовали реализацию алгоритма из работы [15][5].
2. **Алгоритмы на основе LSTM-сетей.** При выборе LSTM-моделей мы опирались на архитектуру трёхслойной LSTM-сети, описанной в работе [17] и реализованной в соответствующем репозитории[6]. Первый слой сети содержит 768 нейронов, а второй и третий — по 512.

[4] https://huggingface.co/datasets/ruscorpora/morphodict-bel
[5] https://github.com/AlexeySorokin/NeuralMorphemeSegmentation
[6] github.com/alesapin/RussianMorphParsing

При этом мы использовали собственную реализацию модели с использованием фреймворка PyTorch [21]. Модель обучалась в течение 55 эпох. В ходе предварительных экспериментов мы приняли решение снизить скорость обучения с оригинальных 10^{-2} до 10^{-3}. Кроме того, мы модифицировали процедуру обучения сети, добавив в неё градиентный клиппинг (в таблице результатов указаны результаты как с клиппингом, так и без него). Наконец, как и в случае свёрточных сетей, мы добавили в эксперимент ансамбль из трёх LSTM-сетей идентичной архитектуры. Наша реализация алгоритма и конфигурационные файлы экспериментов доступны в репозитории https://github.com/ruscorpora/morphemelstm.

3. **Алгоритмы на основе предобученных BERT-подобных моделей.** Мы использовали подход, описанный в работе [19], при котором BERT-подобная модель дообучается в течение 30 эпох на задаче классификации токенов (скорость обучения 4×10^{-6}, размер батча 16). Предварительно токенизированная входная последовательность при этом состоит из букв, составляющих слово. В модификации +lemma во входную последовательность добавляется лемма целиком. Код для обучения и запуска моделей доступен в репозитории https://github.com/ruscorpora/morphemebert. Мы рассмотрели три типа предобученных BERT-подобных моделей:
   (а) белорусскоязычную модель belRoberta-small[7] (16 млн параметров, предобучена на белорусской части набора данных CC-100[8]);
   (б) русскоязычные модели ruBert-base[9] (178 млн параметров, предобучена на 30 ГБ русскоязычных данных) и ruRoberta-large[10] (355 млн параметров, предобучена на 250 ГБ русскоязычных данных) [22];
   (в) многоязычные модели:
      - SlavicBERT[11] (180 млн параметров, обучена на материале четырёх языков: русском, чешском, польском и болгарском) [23];
      - mBERT-base[12] (177 млн параметров, обучена на материале текстов Википедии с использованием 104 наиболее представленных языков) [24];
      - DistilmBERT[13] (134 млн параметров, обучена на том же датасете, что и предыдущая модель) [25].

2.3. Методология

Для сравнения эффективности получившихся моделей мы использовали перекрёстную проверку с пятью подвыборками. Как и в работах [8, 18, 19], мы рассмотрели два подхода к обучению и тестированию моделей, различающихся способом разбиения набора данных на подвыборки. В первом случае разбиение на подвыборки случайное, во втором подвыборки делятся таким образом, чтобы в разных подвыборках не было слов с общим корнем. Второй сценарий разбиения позволяет оценить эффективность алгоритма при работе со словами, содержащими корни, отсутствующие в обучающей выборке. Для оценки качества мы использовали метрики, предложенные в работе [15]: точность (Precision), полноту (Recall) и F-меру для границ морфем, долю верно классифицированных символов (Accuracy) и долю полностью верных разборов (WordAccuracy).

[7] https://huggingface.co/KoichiYasuoka/roberta-small-belarusian
[8] https://data.statmt.org/cc-100/
[9] https://huggingface.co/ai-forever/ruBert-base
[10] https://huggingface.co/ai-forever/ruRoberta-large
[11] https://huggingface.co/DeepPavlov/bert-base-bg-cs-pl-ru-cased
[12] https://huggingface.co/google-bert/bert-base-multilingual-cased
[13] https://huggingface.co/distilbert/distilbert-base-multilingual-cased

Table 2. Experimental results (random split), %

Таблица 2. Результаты экспериментов (случайное разбиение), %

Модель	Точность	Полнота	F-мера	Accuracy	WordAccuracy
1 × CNN	98.18 ± 0.24	97.53 ± 0.26	97.85 ± 0.18	95.93 ± 0.35	87.45 ± 1.03
3 × CNN	98.31 ± 0.16	97.84 ± 0.16	98.08 ± 0.14	96.34 ± 0.26	88.71 ± 0.70
1 × LSTM, без клиппинга	97.95 ± 0.23	98.01 ± 0.16	97.98 ± 0.16	96.14 ± 0.31	88.50 ± 0.95
1 × LSTM, с клиппингом	97.91 ± 0.29	98.19 ± 0.29	98.05 ± 0.18	96.24 ± 0.34	88.80 ± 1.01
3 × LSTM, без клиппинга	98.31 ± 0.14	98.59 ± 0.12	98.45 ± 0.09	97.02 ± 0.18	91.02 ± 0.63
3 × LSTM, с клиппингом	98.42 ± 0.14	98.64 ± 0.22	98.53 ± 0.16	97.17 ± 0.25	91.42 ± 0.81
belRoberta-small	97.94 ± 0.11	98.24 ± 0.16	98.09 ± 0.11	96.26 ± 0.17	88.66 ± 0.53
belRoberta-small+lemma	97.52 ± 0.03	97.89 ± 0.14	97.71 ± 0.09	95.50 ± 0.12	86.80 ± 0.35
ruBert-base	98.32 ± 0.10	98.58 ± 0.14	98.45 ± 0.08	97.02 ± 0.17	90.99 ± 0.44
ruBert-base+lemma	98.08 ± 0.10	98.37 ± 0.12	98.23 ± 0.07	96.58 ± 0.18	89.71 ± 0.39
ruRoberta-large	98.20 ± 0.07	98.38 ± 0.10	98.29 ± 0.07	96.63 ± 0.14	89.64 ± 0.38
ruRoberta-large+lemma	98.19 ± 0.06	98.52 ± 0.08	98.35 ± 0.05	96.75 ± 0.10	89.98 ± 0.34
SlavicBERT	97.18 ± 0.15	97.92 ± 0.16	97.55 ± 0.14	95.17 ± 0.25	85.48 ± 0.61
SlavicBERT+lemma	97.28 ± 0.07	97.83 ± 0.15	97.55 ± 0.10	95.16 ± 0.23	85.40 ± 0.58
mBERT-base	98.40 ± 0.08	98.60 ± 0.15	98.50 ± 0.11	97.11 ± 0.15	91.37 ± 0.39
mBERT-base+lemma	98.23 ± 0.08	98.48 ± 0.11	98.35 ± 0.08	96.80 ± 0.15	90.48 ± 0.36
DistilmBERT	98.30 ± 0.04	98.53 ± 0.16	98.41 ± 0.08	96.90 ± 0.11	90.66 ± 0.24
DistilmBERT+lemma	98.11 ± 0.12	98.41 ± 0.08	98.26 ± 0.08	96.57 ± 0.13	89.76 ± 0.41

3. Результаты и их обсуждение

Полученные результаты в формате **среднее ± среднеквадратичное отклонение** представлены в таблицах 2, 3. В каждом столбце лучший достигнутый результат выделен серым. Все метрики выражены в процентах.

В случае обоих подходов к разбиению наиболее эффективным подходом оказался ансамбль LSTM-сетей с применением градиентного клиппинга при обучении. Доля полностью верных разборов в случае случайного разбиения для этой модели равна 91.42 ± 0.81 %, а в случае разбиения по корням — 73.89 ± 1.06 %. При этом аналогичный ансамбль без клиппинга также оказался в тройке лучших по большинству метрик в обоих случаях, достигнув доли полностью верных разборов 91.02 ± 0.63 % и 73.61 ± 1.16 %, соответственно.

Высокое качество продемонстрировали и дообученные BERT-подобные модели. В случае случайного разбиения высоких результатов удалось добиться при помощи моделей mBERT-base и ruBert-base, обученных без добавления леммы во входную последовательность. Доля полностью верных разборов составила 91.37±0.39 % и 90.99±0.44 %, соответственно. В случае же разбиения по корням лучший результат среди BERT-подобных моделей был достигнут с помощью ruRoberta-large, дообученной с добавлением леммы. В этом случае доля полностью верных разборов оказалась равна 72.83 ± 1.11 %.

Следует отметить, что среди лидеров ни разу не оказалась модель belRoberta-small, обученная на белорусскоязычных данных, при этом русскоязычная ruRoberta-large показала лучший результат среди BERT-подобных моделей в одном из сценариев разбиения. Это показывает, что в задаче морфемной сегментации большие модели, обученные на схожем, но не идентичном домене, могут оказываться более эффективными, чем маленькие, обученные на целевом домене. Кроме того, сравнение трёх многоязычных моделей позволило выделить среди них лидера — mBERT-base, при этом SlavicBERT оказался значительно хуже двух других моделей при случайном разбиении.

Использование ансамблей в случае свёрточных нейронных сетей и LSTM-сетей позволило повысить качество в обоих сценариях обучения и тестирования. Наконец, добавление леммы во входную последовательность BERT-подобных моделей позволило повысить качество для большинства моделей в случае OOV-корней, но при этом заметно снизило качество в случайном разбиении.

Table 3. Experimental results (split by roots), %

Таблица 3. Результаты экспериментов (разбиение по корням), %

Модель	Точность	Полнота	F-мера	Accuracy	WordAccuracy
1 × CNN	94.79 ± 0.36	93.82 ± 0.48	94.30 ± 0.27	89.09 ± 0.48	69.25 ± 1.32
3 × CNN	95.24 ± 0.44	94.11 ± 0.35	94.67 ± 0.34	89.68 ± 0.54	70.84 ± 1.61
1 × LSTM, без клиппинга	94.89 ± 0.47	93.99 ± 0.79	94.44 ± 0.41	89.37 ± 0.78	70.40 ± 1.45
1 × LSTM, с клиппингом	95.17 ± 0.10	93.75 ± 0.66	94.45 ± 0.33	89.31 ± 0.59	70.36 ± 1.59
3 × LSTM, без клиппинга	95.46 ± 0.32	94.91 ± 0.18	95.18 ± 0.23	90.71 ± 0.32	73.61 ± 1.16
3 × LSTM, с клиппингом	95.54 ± 0.30	94.89 ± 0.43	95.21 ± 0.26	90.82 ± 0.42	73.89 ± 1.06
belRoberta-small	95.00 ± 0.27	94.63 ± 0.47	94.82 ± 0.27	89.71 ± 0.58	71.31 ± 1.62
belRoberta-small+lemma	94.74 ± 0.26	94.64 ± 0.28	94.69 ± 0.16	89.67 ± 0.24	71.02 ± 0.95
ruBert-base	95.06 ± 0.26	94.52 ± 0.33	94.79 ± 0.23	89.85 ± 0.44	71.77 ± 1.27
ruBert-base+lemma	94.91 ± 0.35	94.72 ± 0.42	94.82 ± 0.19	89.94 ± 0.31	72.02 ± 1.37
ruRoberta-large	95.21 ± 0.22	94.36 ± 0.36	94.78 ± 0.20	89.61 ± 0.34	71.24 ± 0.99
ruRoberta-large+lemma	95.29 ± 0.28	94.97 ± 0.32	95.13 ± 0.18	90.38 ± 0.38	72.83 ± 1.11
SlavicBERT	94.72 ± 0.32	95.01 ± 0.33	94.87 ± 0.29	89.76 ± 0.58	71.45 ± 1.44
SlavicBERT+lemma	94.68 ± 0.33	94.90 ± 0.30	94.78 ± 0.25	89.64 ± 0.45	71.33 ± 1.33
mBERT-base	95.07 ± 0.30	94.51 ± 0.60	94.79 ± 0.32	89.85 ± 0.59	71.61 ± 1.72
mBERT-base+lemma	95.11 ± 0.28	94.61 ± 0.26	94.86 ± 0.17	89.94 ± 0.37	71.66 ± 1.07
DistilmBERT	95.07 ± 0.36	94.47 ± 0.54	94.76 ± 0.40	89.73 ± 0.77	71.14 ± 2.15
DistilmBERT+lemma	95.13 ± 0.19	94.66 ± 0.26	94.89 ± 0.18	89.95 ± 0.32	71.60 ± 1.26

В сравнении с результатами для белорусского языка, полученными в работе [8], эффективность свёрточных нейронных сетей в наших экспериментах оказалась заметно ниже. Несмотря на преимущество ансамбля CNN над дообученными belRoberta-small и SlavicBERT в случае случайного разбиения, при работе с OOV-корнями свёрточные сети показали результат хуже. Мы связываем это различие с изменившимся составом используемого набора данных, увеличившимся более чем на 10 %, так как в остальном архитектура и реализация моделей были идентичны. В частности, количество уникальных не-корневых морфем выросло почти втрое (с 255 до 751), а среднее число употребления каждой из них при этом снизилось.

Кроме того, в отличие от результатов работы [19], полученных для русского языка, эффективность LSTM-сетей оказалась значительно выше эффективности свёрточных сетей. Мы предполагаем, что это связано в первую очередь не с разницей между русским и белорусским языками, а с различиями в реализации алгоритма и объёмом обучающей выборки.

Следует отметить, что эффективность всех использованных подходов оказалась достаточно близкой. Разница по метрике WordAccuracy между худшим и лучшим алгоритмом составила около 6 % в случае случайного разбиения и около 4.5 % — в случае разбиения по корням. В связи с этим мы приняли решение оценить статистическую значимость превосходства лучшей модели над остальными. При этом прямое применение критериев статистической значимости затруднено малым числом измерений ($N = 5$). Например, согласно T-критерию Вилкоксона для гипотезы «метрика модели А выше, чем модели Б» при $N = 5$ условие «p-value $\geqslant 0.05$» эквивалентно «модель Б оказалась лучше хотя бы на одной из выборок». В связи с этим мы приняли решение дополнить выборку измерений. Для этого мы разбили каждую из тестовых выборок на пять непересекающихся подвыборок и измерили качество работы алгоритмов на каждой из них. Разбиение при этом проходило в той же логике, что и первичное разбиение при кросс-валидации: для случайного разбиения подвыборки выбирались случайно, для разбиения по корням тестовая выборка делилась по корням. После этого мы сопоставили модели с лучшей моделью по метрике WordAccuracy (ансамблем из трёх LSTM-сетей с использованием клиппинга при обучении) с точки зрения T-критерия Вилкоксона. В результате в случае случайного разбиения ансамбль LSTM-сетей с клиппингом статистически значимо лучше всех других алгоритмов, кроме mBERT-base, при сопоставлении с которым p-value

составило 0.21. В случае же разбиения по корням незначительно различаются ансамбли LSTM-сетей с клиппингом и без него (p-value = 0.25), которые значимо превосходят остальные алгоритмы.

При сегментации слов, содержащих OOV-корни, качество работы всех алгоритмов значительно падает. При этом наиболее заметно падение в доле полностью верных разборов (до 20 процентных пунктов). Такое сильное падение именно этой метрики нетрудно объяснить с точки зрения статистики: при доле точно классифицированных символов в 97 % и средней длине слова около 9 символов, можно грубо оценить математическое ожидание доли полностью верных разборов как 0.97^9, то есть $\sim$ 76 %, а при снижении доли точно классифицированных символов до 91 %, доля полностью верных разборов снижается до 0.91^9, то есть $\sim$ 43 %. Разумеется, такая оценка крайне груба, так как классы различных символов в слове не присваиваются независимо друг от друга, а ошибочно разобранные слова в подавляющем большинстве случаев содержат более одного символа с неправильной меткой. Тем не менее, она позволяет приблизительно понять взаимосвязь между использованными метриками.

Помимо подсчёта метрик качества, мы изучили ошибки, допускаемые моделями при сегментации слов с OOV-корнями. Анализ показал, что подавляющее большинство ошибок можно распределить по трём основным типам, причём слова, как правило, содержат не более одной ошибки. Мы выделили следующие типы ошибок:

1. **Сдвиг границы:** в паре верно определённых морфем неправильно определена граница. Среди ошибок этого типа чаще других встречаются ошибки на границе корня и суффикса, например, для *абанемент (абонемент)* сгенерирована сегментация `абанем:ROOT/ент:SUFF` при эталонном `абан:ROOT/емент:SUFF`. Ошибки этого типа также могут встречаться на границе приставки и корня, например, для *ускварыць (ужарить (сало))* сгенерирована сегментация `ус:PREF/квар:ROOT/ы:SUFF/ць:SUFF` при эталонной `у:PREF/сквар:ROOT/ы:SUFF/ць:SUFF`.
2. **Смена типа:** предсказанный разбор отличается от эталонного типом морфем и, возможно, их границами, но количество морфем сохраняется. В рамках этого типа наиболее популярны ошибки с распознаванием пары приставка+корень вместо эталонного корень+суффикс, например, для *насіць (носить)* сгенерирован разбор `на:PREF/сі:ROOT/ць:SUFF` при эталонном `нас:ROOT/і:SUFF/ць:SUFF`; а также с распознаванием пары корень+суффикс вместо приставка+корень, например, для *абдаць (обдать)* сгенерирован разбор `абд:ROOT/а:SUFF/ць:SUFF` при эталонном `аб:PREF/да:ROOT/ць:SUFF`.
3. **Смена количества:** предсказанный разбор отличается от эталонного числом морфем. Среди ошибок этого типа распространены ситуации со склеиванием корня и суффикса или приставки и корня в единственную морфему. Примером такой ошибки является разбор `бів:ROOT/ень:SUFF` при эталонном `бі:ROOT/в:SUFF/ень:SUFF` (для *бівень (бивень)*). Вторым частотным случаем является вычленение избыточной приставки или суффикса из корня, например, слово *спешна (спешно)* сегментировано как `с:PREF/пеш:ROOT/н:SUFF/а:SUFF` при эталонном `спеш:ROOT/н:SUFF/а:SUFF`.

Распространённость перечисленных типов ошибок в зависимости от модели приведена в таблице 4. Сопоставив результаты для разных моделей, можно заметить, что модели на базе свёрточных нейронных сетей отличаются от прочих моделей большим уклоном в избыточную сегментацию морфем и меньшим процентом случаев склеивания. Распределение ошибок для прочих моделей можно считать схожим.

Важно отметить, что для белорусского языка характер ошибок схож с результатами, ранее полученными для прочих славянских языков [8, 19]. Большинство ошибок связано с определением границ корня, тогда как другие варианты ошибок, например, определение неправильной метки для морфемы, встретились значительно реже.

Table 4. Percentage of the most frequent error types among the model's errors, %. Error types: 1A — root-suffix boundary shift, 1B — prefix-root boundary shift, 2A — replacement of a root+suffix morpheme pair with a prefix+root pair, 2B — replacement of a prefix+root morpheme pair with a root+suffix pair, 3A — splitting a root into a prefix+root morpheme pair, 3B — splitting a root into a root+suffix morpheme pair, 3C — merging a prefix+root morpheme pair into a single root morpheme, 3D — merging a root+suffix morpheme pair into a single root morpheme.

Таблица 4. Доля наиболее популярных типов ошибок среди ошибок модели, %. Типы ошибок: 1A — сдвиг границы корень-суффикс, 1B — сдвиг границы приставка-корень, 2A — замена пары морфем корень-суффикс на пару приставка-корень, 2B — замена пары морфем приставка-корень на пару корень-суффикс, 3A — разбиение корня на пару морфем приставка+корень, 3B — разбиение корня на пару морфем корень+суффикс, 3C — склеивание пары морфем приставка+корень в одну морфему, 3D — склеивание пары морфем корень+суффикс в одну морфему.

Модель	1A	1B	2A	2B	3A	3B	3C	3D
CNN, 1 модель	16.5	2.9	3.4	2.9	22.4	16.4	6.4	11.9
CNN, 3 модели	13.8	2.1	4.2	2.4	24.0	18.8	5.2	10.9
LSTM, 1 модель, без клиппинга	16.4	1.9	3.8	2.2	14.1	14.3	13.7	16.6
LSTM, 1 модель, с клиппингом	14.9	2.9	3.5	1.8	12.6	13.3	15.5	18.2
LSTM, 3 модели, без клиппинга	13.8	1.9	3.0	1.9	16.2	14.8	13.8	17.9
LSTM, 3 модели, с клиппингом	16.9	< 1	3.6	1.7	17.2	15.9	10.3	17.4
belRoberta-small	14.1	2.2	2.9	1.6	14.5	15.6	11.9	14.9
belRoberta-small+lemma	14.8	1.6	2.6	1.8	16.5	15.4	11.8	15.4
ruBert-base	14.6	1.8	3.3	2.9	15.1	14.2	13.4	15.1
ruBert-base+lemma	16.9	< 1	3.8	1.7	17.6	15.5	8.6	14.4
ruRoberta-large	13.7	1.6	3.3	1.9	14.3	14.7	12.2	17.9
ruRoberta-large+lemma	13.1	< 1	3.3	< 1	16.6	14.9	12.1	18.1
SlavicBERT	14.1	1.8	3.3	1.6	17.9	15.5	8.1	14.6
SlavicBERT+lemma	14.8	1.9	4.4	1.6	17.8	14.8	8.5	15.1
mBERT-base	14.1	1.7	4.5	2.3	15.2	15.2	11.1	14.6
mBERT-base+lemma	13.7	2.1	3.6	< 1	17.4	12.8	14.9	16.5
DistilmBERT	15.0	1.8	3.1	< 1	15.6	14.4	14.1	14.3
DistilmBERT+lemma	12.5	1.6	3.4	< 1	17.4	14.9	11.0	17.3

Заключение

В настоящей работе впервые проведено масштабное экспериментальное сравнение актуальных алгоритмов морфемной сегментации на материале белорусского языка. Были протестированы три основных подхода: с использованием свёрточных нейронных сетей, с использованием LSTM-сетей и с дообучением BERT-подобных моделей. Из-за нехватки моноязычных белорусских языковых моделей, эксперименты проводились с белорусскоязычными, русскоязычными и многоязычными моделями. Тестирование проводилось на свободно распространяемом наборе данных Slounik с использованием двух стратегий разбиения данных на обучающую и тестовую выборки. В первом случае разбиение было случайным, во втором случае использовалось разбиение по корням: однокоренные слова не могли попасть одновременно и в обучающую, и в тестовую выборки. По итогам экспериментов лучший результат в случае обеих стратегий показал ансамбль из трёх трёхслойных LSTM-сетей, обученных с использованием градиентного клиппинга. Доля полностью верных разборов при случайном разбиении составила 91.42 %, при разбиении по корням — 73.89 %. Веса моделей обученного на случайном разбиении ансамбля доступны в репозитории с реализацией алгоритма[14]. Близкие результаты продемонстрировали дообученные BERT-подобные модели: в случае случайного разбиения лучшие результаты достигнуты при помощи русскоязычной ruBert-base и многоязычной mBERT-base, в случае разбиения по корням — при помощи русскоязычной ruRoberta-large.

[14] https://github.com/ruscorpora/morphemelstm

При этом сравнительно слабые результаты белорусскоязычной модели belRoberta-small, вероятнее всего, связаны с её крайне малым размером. В ближайшем будущем мы планируем опубликовать лучшие из дообученных моделей на HuggingFace[15].

Следует заметить, что в работе рассматривалась исключительно сегментация лемм слов, кроме того, игнорировалась потенциальная омонимия. Переход от лемм к словоформам, а также учёт контекста конкретного словоупотребления может быть востребован в задачах построения токенизаторов и должен быть изучен подробнее в будущем. К прочим ограничениям нашего исследования следует отнести отсутствие среди рассмотренных подходов, опирающихся на большие языковые модели. В дальнейшем мы планируем проанализировать возможность использования LLM для повышения качества распознавания корней. Кроме того, мы проведём аналогичные эксперименты для других малоресурсных языков, в том числе, славянских, и обобщим полученные данные об эффективности различных подходов к морфемной сегментации в ситуации малого объёма аннотированных обучающих данных и отсутствия крупных предобученных моделей для целевого языка.

References

[1] P. Gage, "A new algorithm for data compression", *C User Journal*, vol. 12, no. 2, pp. 23–38, 1994.

[2] A. Matthews, G. Neubig, and C. Dyer, "Using morphological knowledge in open-vocabulary neural language models", in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, 2018, pp. 1435–1445. DOI: 10.18653/v1/N18-1130.

[3] A. Nzeyimana and A. Niyongabo Rubungo, "KinyaBERT: A morphology-aware Kinyarwanda language model", in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 5347–5363. DOI: 10.18653/v1/2022.acl-long.367.

[4] M. W. Kildeberg, E. A. Schledermann, N. Larsen, and R. van der Goot, *From smør-re-brød to subwords: Training LLMs on Danish, one morpheme at a time*, 2025. arXiv: 2504.01540 [cs.CL].

[5] E. Asgari, Y. E. Kheir, and M. A. S. Javaheri, *MorphBPE: A morpho-aware tokenizer bridging linguistic complexity for efficient LLM training across morphologies*, 2025. arXiv: 2502.00894 [cs.CL].

[6] S. O. Savchuk *et al.*, "Russian National Corpus 2.0: New opportunities and development prospects", *Voprosy Jazykoznanija*, no. 2, pp. 7–34, 2024, in Russian. DOI: 10.31857/0373-658x.2024.2.7-34.

[7] M. Olbrich and Z. Žabokrtský, "Morphological segmentation with neural networks: Performance effects of architecture, data size, and cross-lingual transfer in seven languages", in *Text, Speech, and Dialogue*, 2026, pp. 275–286. DOI: 10.1007/978-3-032-02551-7_24.

[8] D. Morozov, L. Astapenka, A. Glazkova, T. Garipov, and O. Lyashevskaya, "BERT-like models for Slavic morpheme segmentation", in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 6795–6815. DOI: 10.18653/v1/2025.acl-long.337.

[9] M. Pranjić, M. Robnik-Šikonja, and S. Pollak, "LLMSegm: Surface-level morphological segmentation using large language model", in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 10 665–10 674.

[10] C. Anderson, M. Nguyen, and R. Coto-Solano, "Unsupervised, semi-supervised and LLM-based morphological segmentation for Bribri", in *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, 2025, pp. 63–76. DOI: 10.18653/v1/2025.americasnlp-1.7.

[15] https://huggingface.co/ruscorpora/models

[11] K. Batsuren *et al.*, "The SIGMORPHON 2022 shared task on morpheme segmentation", in *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 2022, pp. 103–116. DOI: 10.18653/v1/2022.sigmorphon-1.11.

[12] B. Peters and A. F. T. Martins, "Beyond characters: Subword-level morpheme segmentation", in *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 2022, pp. 131–138. DOI: 10.18653/v1/2022.sigmorphon-1.14.

[13] S. Wehrli, S. Clematide, and P. Makarov, "CLUZH at SIGMORPHON 2022 shared tasks on morpheme segmentation and inflection generation", in *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 2022, pp. 212–219. DOI: 10.18653/v1/2022.sigmorphon-1.21.

[14] A. Sorokin, "Improving morpheme segmentation using BERT embeddings", in *Analysis of Images, Social Networks and Texts*, 2022, pp. 148–161. DOI: 10.1007/978-3-031-16500-9_13.

[15] A. Sorokin and A. Kravtsova, "Deep convolutional networks for supervised morpheme segmentation of Russian language", in *Artificial Intelligence and Natural Language*, 2018, pp. 3–10. DOI: 10.1007/978-3-030-01204-5_1.

[16] E. I. Bolshakova and A. S. Sapin, "Comparing models of morpheme analysis for Russian words based on machine learning", in *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"*, vol. 18, 2019, pp. 104–113.

[17] E. Bolshakova and A. Sapin, "Bi-LSTM model for morpheme segmentation of Russian words", in *Artificial Intelligence and Natural Language*, 2019, pp. 151–160. DOI: 10.1007/978-3-030-34518-1_11.

[18] T. Garipov, D. Morozov, and A. Glazkova, "Generalization ability of CNN-based Morpheme Segmentation", in *Proceedings of the Ivannikov Ispras Open Conference (ISPRAS)*, 2024, pp. 58–62. DOI: 10.1109/ISPRAS60948.2023.10508171.

[19] D. Morozov, T. Garipov, O. Lyashevskaya, S. Savchuk, B. Iomdin, and A. Glazkova, "Automatic morpheme segmentation for Russian: Can an algorithm replace experts?", *Journal of Language and Education*, vol. 10, no. 4, pp. 71–84, 2024. DOI: 10.17323/jle.2024.22237.

[20] L. S. Mormysh, A. M. Bardovich, and L. M. Shakun, *School morpheme dictionary of the Belarusian language [SHkol'ny marfemny slovnik belaruskaj movy]*. Minsk: Aversev, 2005, in Belarusian.

[21] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library", in *Advances in Neural Information Processing Systems 32*, 2019, pp. 8024–8035.

[22] D. Zmitrovich *et al.*, "A family of pretrained transformer language models for Russian", in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 507–524.

[23] M. Arkhipov, M. Trofimova, Y. Kuratov, and A. Sorokin, "Tuning multilingual transformers for language-specific named entity recognition", in *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, 2019, pp. 89–93. DOI: 10.18653/v1/W19-3712.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[25] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*, 2020. arXiv: 1910.01108 [cs.CL].