https://doi.org/10.17323/jle.2024.22237

Автоматическое построение морфемных разборов слов русского языка: Могут ли алгоритмические подходы заменить экспертов?

Морозов Дмитрий Алексеевич [1], **Гарипов Тимур Александрович** [1], **Ляшевская Ольга Николаевна** [2,3], **Савчук Светлана Олеговна** [3], **Иомдин Борис Леонидович** [4], **Глазкова Анна Валерьевна** [5]

[1] Новосибирский государственный университет, г. Новосибирск, Россия

[2] Национальный исследовательский унивеоВШЭ, Москва, Россия

[3] Институт русского языка имени В. В. Виноградова РАН, г. Москва, Россия

[4] Независимый исследователь

[5] Тюменский государственный университет, г. Тюмень, Россия

АННОТАЦИЯ

Введение: Несмотря на интерес исследователей к алгоритмам автоматического построения морфемных разборов слов русского языка, различия в постановке задачи и использованных наборах данных затрудняют сравнение полученных разными авторами результатов. Более того, неясно, связаны ли ошибки моделей с принципиальной нехваткой обобщающей способности алгоритмов или же с ошибками и непоследовательной разметкой в обучающей выборке. Остаётся неясным, может ли какой-либо из представленных алгоритмов использоваться для автоматического пополнения существующих морфемных словарей.

Цель: Сравнение алгоритмов автоматического построения морфемных разборов для слов русского языка с точки зрения применимости этих алгоритмов для автоматического расширения существующих морфемных словарей.

Методология: Проведено сравнение нескольких актуальных алгоритмов, опирающихся на методы машинного обучения, на материале трёх наборов данных, размеченных в разной парадигме морфемного членения. Были реализованы две серии экспериментов, в каждой из которых качество алгоритма измерялось в ходе пятикратной перекрёстной проверки. В первом случае набор данных разделялся на пять непересекающихся подвыборок случайным образом. Во втором случае набор реализовывался таким образом, чтобы все слова, содержащие некоторый корень, обязательно попали в одну и ту же подвыборку, при этом слова с несколькими корнями были заранее исключены из рассмотрения. В ходе перекрёстной проверки мы обучали модель на совокупности четырёх подвыборок и оценивали итоговое качество на пятой.

Результаты: В обоих случаях наибольшую эффективность показали алгоритмы, опирающиеся на ансамбль свёрточных нейронных сетей. Обнаружено, что качество разметки моделью в первой серии экспериментов практически достигает уровня экспертов. При этом зарегистрировано значительное снижение качества порождаемой разметки во второй серии экспериментов относительно первой, то есть при работе со словами, содержащими корни, не встретившиеся модели при обучении.

Заключение: Хотя в среднем качество автоматической разметки достигло уровня, близкого к экспертному, отсутствующий или недостаточный учёт семантики в таких моделях делает невозможным автоматическое расширение словарей без валидации результатов разметки экспертами. Дальнейшим направлением развития в этой области должно стать преодоление ключевой проблемы всех исследованных алгоритмов: низкого качества разметки при работе с сокращениями и корнями, неизвестными модели. Если допустимо предположить, что в составе тестовой выборки множество корней мало отличимо от аналогичного множества в составе обучающей выборки, то для разметки следует использовать алгоритм на основе ансамбля свёрточных нейронных сетей, показавший в ходе исследования лучшее качество.

КЛЮЧЕВЫЕ СЛОВА

автоматическая сегментация морфем, морфология русского языка, расширение словаря, морфологический анализ, обработка естественного языка, производительность экспертного уровня, сверточные нейронные сети

Для цитирования: Морозов, Д.А., Гарипов, Т.А., Ляшевская, О.Н., Савчук, С.О., Иомдин, Б.Л., & Глазкова, А.В. (2024). Автоматическая сегментация морфем для русского языка: Может ли алгоритм заменить экспертов? *Journal of Language and Education, 10*(4), 74-87. https://doi.org/10.17323/jle.2024.22237

Correspondence:
Морозов Дмитрий Алексеевич,
morozowdm@gmail.com

Получена: 18 августа 2024

Принята: 16 декабря 2024

Опубликована: 30 декабря 2024

CC BY

ВВЕДЕНИЕ

Морфемный разбор слова представляет собой процесс разбиения слова на минимальные неделимые с точки зрения языка единицы — морфемы, к которым относят, например, приставки, корни и суффиксы. Многие из правил орфографии и правописания, изучаемых в школе, опираются на умение ученика выделить морфему или определить взаимное расположение нескольких морфем (Бакулина, 2012). В русском языке к таким правилам относятся, например, написание глухих и звонких согласных в приставках, выбор между *-н-* и -нн- на границах морфем и внутри них, проверка написания безударных гласных в корне при помощи поиска однокоренных слов.

Алгоритмы морфемной сегментации могут быть использованы и при разработке инструментов обработки естественного языка, как для обогащения признакового описания текста, например, в задаче автоматической оценки сложности текста (Морозов и др., 2024), так и при создании токенизаторов для языковых моделей. Предыдущие исследования показали, что использование морфемных токенизаторов в качестве альтернативы общепринятым Byte-Pair Encoding (BPE) токенизаторам может улучшить качество итоговой модели (Matthews et al., 2018).

При этом доля слов, не описанных в морфемных словарях, остается значительной. Так, в одном из самых больших словарей для русского языка — «Словообразовательном словаре русского языка» (Тихонов, 1990) — содержатся разборы около 150 тысяч лемм, тогда как в Основном корпусе Национального корпуса русского языка (Савчук и др., 2024) можно найти более 250 тысяч уникальных лемм. Таким образом, разработка алгоритмов для автоматической «доразметки» слов представляет собой востребованную задачу, которая в случае русского языка осложняется отсутствием единого подхода к сегментации слова на морфемы (Иомдин, 2019).

Некоторые составители словарей опираются в своей работе на так называемый критерий Винокура (Винокур, 1946). В этом случае, чтобы выделить морфему, необходимо восстановить словообразовательную цепочку, то есть найти такое исходное слово, которое при добавлении морфемы, даст искомое, например, пис-а-ть + -тель = пис-а-тель. Этот подход, в частности, использовался при составлении упомянутого выше «Словообразовательного словаря», а также он часто применяется в рамках курса школьного образования. Существенным недостатком такого подхода является то, что слова, воспринимаемые носителями языка как однокоренные, таковыми не являются с точки зрения словаря. Так, корнем слова неодобрительный, согласно «Словообразовательному словарю», является строка -одобр-, а в добро — -добр-, что делает эти слова формально неоднокоренными.

Другие лексикографы, такие как авторы «Словаря морфем русского языка» (Кузнецова, Ефремова, 1986), выбирают более детальный подход, основанный на сопоставлении слова с другими лексемами схожей структуры. Например, корнем слова улыбаться является -лыб-, так как структура этого слова аналогична структуре других глаголов на у- (у-смех-а-ть-ся). В некоторых случаях слова анализируют с учетом этимологии, даже если она не является прозрачной для носителей (на-сек-ом-ое, вос-точ-н-ый). В иностранных словах заимствованные основы разбиваются (например, ре-волюц-и-я, квит-анци-я), если усматривается семантическое и структурное соответствие между ними и лексемами похожего строения (ср. э-волюц-и-я, рас-квит-а-ть-ся).

Кроме того, подробное изучение словарей позволяет заметить, что в отдельных случаях авторы склонны принимать решения, противоречащие заявленной исходно парадигме. Таким образом, правила морфемной сегментации представляют собой достаточно слабо формализованную область, что, по-видимому, делает невозможным создание идеального, не допускающего ошибок автоматического алгоритма.

В то же время, учитывая высокий прикладной потенциал задачи, к настоящему времени создано множество приближенных решений. Одним из наиболее часто упоминаемых и хорошо изученных является семейство подходов, опирающихся на алгоритм Morfessor (Creutz & Lagus, 2002). В состав этого семейства входят алгоритмы обучения без учителя и с частичным привлечением учителя. Большинство алгоритмов этого семейства могут быть обучены на материале любого языка. Среди наиболее значимых модификаций оригинального алгоритма стоит упомянуть подход S.-A. Grönroos et al. (2020), в котором исследуется комбинация Morfessor с алгоритмом EM+Prune.

Значительный прогресс в качестве автоматической разметки был достигнут в ходе соревнования SIGMORPHON-2022 (Batsuren et al., 2022), где были представлены несколько подходов, существенно превзошедших базовые подходы, включая Morfessor, ULM (Kudo, 2018) и WordPiece (Schuster & Nakajima, 2012). Среди предложенных архитектур можно выделить модели на основе Transformer (Zundi & Avaajargal, 2022; Peters & Martins, 2022), GRU-модели (Levine, 2022), hard-attention трансдьюсеры (Wehrlie et al., 2022), нейронные сети LSTM (Peters & Martins, 2022; Girrbach, 2022) и скрытые марковские модели (Bodnár, 2022). Команда DeepSPIN (Peters & Martins, 2022) достигла лучшего качества разметки по всем девяти языкам, представленным на соревновании. Их решения были основаны на использовании сетей LSTM с особой функцией потерь (DeepSPIN-1 и DeepSPIN-2) и архитектуры Transformer (DeepSPIN-3).

В ближайшем будущем следует ожидать быстрого роста числа подходов, опирающихся на большие языковые модели. Так, Pranjić et al. (2024) предложили алгоритм, основанный на сети Glot500-m (ImaniGooghari et al., 2023), который представляет собой бинарный классификатор для поиска границ морфем в слове. В то же время недостатки этого алгоритма в настоящее время не позволяют считать этот

подход конкурентным. К таковым следует отнести сравнительно невысокое качество для английского, финского и турецкого языков, а также длительное время разметки (алгоритм последовательно проверяет каждую пару соседних букв в слове).

Для русского языка актуальные решения, превосходящие алгоритм Morfessor, представлены в работах А. Сорокина и А. Кравцовой (2018), а также Е. Большаковой и А. Сапина (2019a; 2019b). Авторы рассматривают подходы, основанные на свёрточных нейронных сетях, сетях LSTM и градиентном бустинге над деревьями решений. Результаты сравнения этих алгоритмов на двух различных наборах данных не позволяют сделать однозначный вывод о превосходстве какого-либо одного алгоритма. Тем не менее, достигнутое качество разметки (около 90 % полностью корректных разборов) является достаточно высоким. Русский язык также был включен в число языков на соревновании SIGMORPHON-2022, где наилучшее качество показала модель DeepSPIN-3.

В то же время ряд вопросов в этой области остается недостаточно изученным. Garipov et al. (2023) обнаружили, что модель, основанная на сверточных нейронных сетях, имеет существенный недостаток: её качество резко снижается при тестировании на словах, содержащих корни, которых не было в обучающей выборке, при этом процент полностью корректных разборов падает на 17–18 %. Остается неясным, существует ли аналогичная проблема для других алгоритмов, демонстрирующих высокое качество. Кроме того, в большинстве случаев авторы используют в экспериментах только один морфемный словарь, тогда как для получения объективных результатов имеет смысл учитывать большее количество словарей.

Сравнение алгоритмов дополнительно осложняется тем, что исследователи фактически решают разные задачи. В некоторых случаях (Sorokin & Kravtsova, 2018; Bolshakova & Sapin, 2019a; Bolshakova & Sapin, 2019b) рассматривается задача «поверхностной» сегментации, заключающаяся исключительно в разбиении строки на подстроки, в то время как на соревновании SIGMORPHON рассматривалась задача, включающая в себя восстановление исходной формы морфем, — так называемая каноническая сегментация. В качестве иллюстрации разницы между этими подходами Cotterell et al. (2016) приводит следующий пример: поверхностная сегментация английского слова funniest (самый смешной) будет равна funn-i-est, то есть конкатенация получившихся морфем в точности равна исходному слову, тогда как при «канонической» сегментации будет произведен разбор fun-y-est.

Еще одной проблемой является отделение ошибок разметки, связанных с внутренней противоречивостью исходных словарей. Доля таких ошибок может быть оценена при помощи сравнения словарной и экспертной разметки, однако ранее такие эксперименты для русского языка не проводились.

Таким образом, целью настоящего исследования является сравнение разнообразных существующих алгоритмов морфемного анализа слов русского языка и оценка их применимости для автоматического расширения существующих морфемных словарей. В работе мы ставим перед собой следующие вопросы (RQ):

RQ1: Какой из представленных ранее алгоритмов достигает лучшего качества разметки на материале различных морфемных словарей русского языка, опирающихся на различные парадигмы построения морфемных разборов?
RQ2: Насколько эффективно существующие алгоритмы способны справляться с анализом слов, содержащих корни, не представленные в обучающей выборке?
RQ3: Как соотносится качество автоматической сегментации и экспертной разметки?

МЕТОДОЛОГИЯ

Данные

В ходе исследования мы использовали морфемные словари, в которых каждое слово сегментировано на морфемы с указанием типа каждой из них. Мы рассматривали морфемы семи типов: PREF (приставка), ROOT (корень), SUFF (суффикс), END (окончание), POST (постфикс), LINK (соединительная гласная), HYPH (дефис). Для обеспечения большей репрезентативности нашего исследования мы использовали три набора данных, различающихся парадигмами морфемного членения:

(1) Morphodict-K: набор данных, основанный на «Словаре морфем русского языка» (Кузнецова, Ефремова, 1986) и используемый в Основном корпусе Национального корпуса русского языка. Правила сегментации в этом наборе данных подразумевают значительную (хотя и не максимальную) дробность выделения морфем и соотносимость с другими лексемами аналогичного строения.

(2) Morphodict-T: набор данных, основанный на «Словообразовательном словаре русского языка» (Тихонов, 1990) и используемый в Обучающем корпусе Национального корпуса русского языка. Для разделения на морфемы в нем используется так называемый критерий Винокура. Морфемы в Morphodict-T менее дробны, чем в Morphodict-K. Так, Morphodict-T содержит разборы улыб-а-ть-ся, насеком-ое, восточ-н-ый, а Morphodict-K — у-лыб-а-ть-ся, на-сек-ом-ое, вос-точ-н-ый. Особенно расхождения заметны в сегментации заимствований, например, революци-я и квитанци-я в Morphodict-T, революц-и-я и квит-анци-я в Morphodict-K. Различается и

словарный состав наборов. Из 75 649 слов, разобранных в Morphodict-K, немногим более 58 тысяч содержатся в Morphodict-T. Следует отметить, что использованный нами набор данных Morphodict-T отличается от аналогичного датасета, использованного А. Сорокиным и А. Кравцовой (2018), благодаря значительной корректировке ошибок в разметке типов морфем. Поиск ошибок и их исправление были проведены тремя экспертами. В тех случаях, когда предлагаемые экспертами исправления не совпадали, соответствующий морфемный разбор удалялся из дальнейшего рассмотрения (всего 27 случаев). Всего экспертами было внесены правки в 31 468 разборов.

(3) CrossLexica (Большаков, 2013) — это набор данных, использованный в работах Bolshakova & Sapin (2019a), Bolshakova & Sapin (2019b). Правила морфемной сегментации для этого набора данных не описаны в явном виде, однако в этом датасете встречаются расхождения как с разборами из Morphodict-K, так и с разборами из Morphodict-T (Таблица 1). В отличие от двух предыдущих наборов данных, в датасете CrossLexica отсутствуют слова с несколькими корнями, но присутствует небольшое количество нелемматизированных слов.

Краткое описание набора данных приведено в Таблице 2.

Важно отметить, что в рамках исследования мы считали, что слово в точности равно конкатенации составляющих его морфем, что в общем случае неверно. Так, слово горбунья может быть разобрано как горб:ROOT/ун:SUFF/ьj:SUFF/я:END (Кузнецова, Ефремова, 1986) с добавлением *-j-*, отсутствующим в исходном слове. В таких случаях мы модифицировали сегментацию, исключая добавленный *-j-*. Если после этого *-ь-* оставался единственной буквой в морфеме, мы объединяли эту морфему с предыдущей. Таким образом, в рассмотренном случае упрощенная сегментация, использовавшаяся в экспериментах, выглядела как горб:ROOT/унь:SUFF/я:END.

Другой важной особенностью нашей работы является то, что в основном рассматривались именно леммы. Это ограничивает область потенциального применения моделей, обученных в ходе экспериментов, в то же время это позволяет считать несущественной проблему омонимичных слов с различными разборами, так как подобные случаи достаточно редко встречаются среди лемм русского языка.

Алгоритмы

Алгоритмы с определением типа морфем

Среди алгоритмов, решающих помимо задачи сегментации и задачу определения типа морфем, мы рассмотрели три, показавшие лучшее качество в предыдущих экспериментах: ансамбль свёрточных нейронных сетей (далее CNN) (Sorokin & Kravtsova, 2018), алгоритм градиентного бустинга над деревьями решений (далее GBDT) и алгоритм, опирающийся на LSTM-сеть (далее LSTM). Сравнение этих алгоритмов ранее не позволило установить однозначного преимущества ни одного из подходов (Bolshakova & Sapin, 2019a; Bolshakova & Sapin, 2019b). Для проведения более полного и объективного сравнения мы решили воспроизвести обучение и оценку качества, используя данные из трех перечисленных наборов. Небольшим дополнительным аспектом исследования стало использование морфологических характеристик слов для улучшения качества разметки при помощи алгоритмов GBDT и LSTM, предложенное Е.

Таблица 1

Примеры различий в разметке между наборами данных

Слово	Morphodict-K	Morphodict-T	CrossLexica
революция ‘revolution’	ре:PREF/волюц:ROOT/и:SUFF/я:END	революци:ROOT/я:END	ре:PREF/вол:ROOT/юци:SUFF/я:END
утверждать ‘to approve’	у:PREF/тверждд:ROOT/а:SUFF/ть:END	утвержд:ROOT/а:SUFF/ть:END	у:PREF/тверждд:ROOT/ать:END
собственник ‘owner’	соб:ROOT/ств:SUFF/енн:SUFF/ик:SUFF	собственн:ROOT/ик:SUFF	соб:ROOT/ств:SUFF/ен:SUFF/ник:SUFF

Таблица 2

Некоторые характеристики использованных наборов данных

Характеристика	CrossLexica	Morphodict-T	Morphodict-K
Количество уникальных слов	23426	95895	75649
Количество уникальных морфем	2745	15899	8079
Количество уникальных корней	2256	15253	7148
Среднее число морфем в слове	3.68	3.86	4.12
Среднее количество употреблений морфемы в наборе данных	25.14	23.29	38.56
Среднее количество употреблений корня в наборе данных	8.31	7.54	12.24
Средняя длина корня	4.57	5.52	4.62

Большаковой и А. Сапиным (2019a; 2019b). Мы поставили задачу изучить влияние морфологических признаков на качество работы этих алгоритмов.

Таким образом, нами были исследовали три алгоритма построения морфемных разборов с определением типа морфем:

(1) CNN. Мы использовали реализацию из оригинальной статьи[1]. Модель представляет собой ансамбль из трех идентичных по архитектуре трехслойных свёрточных нейронных сетей с размером окна 5 и 192 фильтрами. Мы обучали модели в течение 25 эпох с критерием ранней остановки — 10 последовательных эпох без повышения качества при валидации.
(2) LSTM. Мы использовали реализацию алгоритма из репозитория[2] без каких-либо изменений.
(3) GBDT. Мы использовали реализацию алгоритма из репозитория[3] без каких-либо изменений.

К сожалению, в репозиториях не были указаны необходимые версии библиотек, в связи с чем нам пришлось подобрать произвольные версии, не нарушающие совместимость использованных библиотек.

Каждый из перечисленных алгоритмов является побуквенным классификатором. Каждому символу (букве) исходного слова с помощью алгоритма присваивается двусоставная метка. Первая часть метки отвечает за положение символа внутри морфемы: B соответствует первой букве морфемы, M соответствует средней (не первой и не последней) букве, E соответствует последней букве морфемы, S присваивается буквам, составляющим всю морфему целиком (например, дефисам всегда присваивается метка S). Вторая часть морфемы отвечает за тип морфемы. Например, слову *слово* с морфемным разбором слов:*ROOT*/о:*END* будет соответствовать последовательность меток: B-ROOT, M-ROOT, M-ROOT, E-ROOT, S-END.

Алгоритмы без определения типа морфем

Большинство алгоритмов построения морфемных разборов, показавших высокое качество для русского языка, являются алгоритмами с определением типа морфем. Тем не менее, на соревновании SIGMORPHON в 2022 г. (Batsuren et al., 2022) задача построения «канонического» разбора без определения типа морфем рассматривалась для девяти языков, включая русский. Лучшие результаты как для русского, так и для других языков, продемонстрировали модели команды DeepSPIN (Peters & Martins, 2022), причем достигнутое качество разметки оказалось чрезвычайно высоким. В то же время использованный на SIGMORPHON набор данных во многом состоял из словоформ, а не лемм, что могло существенно повлиять на качество генерируемой разметки, особенно в тех случаях, когда в обучающей и тестовой выборках попадаются формы одного и того же слова. Важной особенностью датасета являлось то, что «канонический» разбор скорее являлся словообразовательной цепочкой, нежели набором морфем. Так, слову предугадывавшую сопоставлялся псевдоразбор «пред @@у @@гадать @@ывать @@вший @@ую». Существенные различия в постановке эксперимента и наборе данных не позволяют напрямую соотнести результаты, полученные в ходе соревнования, с прочими работами. В связи с этим мы решили проверить качество разметки наших данных с помощи модели DeepSPIN-3, основанной на архитектуре Transformer и функционирующей на уровне фрагментов слов.

Дополнительно мы рассмотрели модель, представленную А. Сорокиным (2022) в качестве усовершенствования предыдущего подхода (Sorokin & Kravtsova, 2018). В ней помимо символьных n-грамм в качестве признакового описания используются векторные представления слов, получаемые при помощи BERT-подобного кодировщика. Прямое сравнение улучшенной модели невозможно, так как в реализации алгоритма отсутствует определение типа морфемы, а в оригинальной работе не представлен русский язык. Чтобы оценить влияние дополнительной информации на качество итоговой разметки, мы обучили две модели: ансамбль CNN без определения типа морфем, использующий только символьные n-граммы, и аналогичный ансамбль, но с расширенным признаковым описанием. Наша гипотеза состояла в том, что использование предобученных векторных моделей позволит учесть семантику слов, что способствует предотвращению снижения качества на словах, содержащих незнакомые модели корни (Garipov et al., 2023). Обе модели решали задачу посимвольной классификации, как и в случае алгоритмов с определением типа морфем.

Таким образом, мы рассмотрели три алгоритма построения морфемных разборов без определения типа морфем:

(1) DeepSPIN-3. Мы использовали реализацию из репозитория[4]. Размер словаря модели был выбран равным 4000 в связи с небольшим размером наборов данных. Остальные гиперпараметры модели были выбраны в соответствии с настройками, представленными в оригинальной статье.
(2) TorchCNN. Ансамбль свёрточных нейронных сетей без определения типа морфем. Мы использовали реализацию из репозитория[5].

[1] NeuralMorphemeSegmentation (Библиотека Python). A. Sorokin. https://github.com/AlexeySorokin/NeuralMorphemeSegmentation
[2] RussianMorphParsing (Библиотека Python). A. Sapin. https://github.com/alesapin/RussianMorphParsing
[3] RussianMorphParsing (Библиотека Python). A. Sapin. https://github.com/alesapin/RussianMorphParsing
[4] MorphemeSegmentation (Библиотека Python). J. Stephenson. https://github.com/joshstephenson/MorphemeSegmentation
[5] MorphemeBert (Библиотека Python). A. Sorokin. https://github.com/AlexeySorokin/MorphemeBert

(3) MorphemeBERT. Ансамбль свёрточных нейронных сетей без определения типа морфем, использующий в качестве дополнительной информации векторные представления из BERT-подобной модели. Мы использовали реализацию из репозитория[6]. В качестве источника векторных представлений использовалась модель rubert-base-cased (Kuratov & Arkhipov, 2019).

Эксперименты

Эксперименты в рамках RQ1

Для того чтобы ответить на RQ1, мы последовательно обучили все модели на всех доступных наборах данных и измерили качество их разметки с помощью кросс-валидации с разбиением на пять подвыборок. Для алгоритмов GBDT и LSTM было обучено по три варианта моделей: 1) без использования дополнительной информации; 2) с использованием частей речи и лемм; 3) с использованием лемм и всей имеющейся морфологической информации, генерируемой PyMorphy2.

Эксперименты в рамках RQ2

Для того чтобы ответить на RQ2, мы разделили все корни, присутствующие в наборе данных, на пять равных подмножеств, а затем сопоставили каждому подмножеству корней подмножество слов, содержащих эти корни. Слова несколькими корнями были исключены из набора данных заранее. Затем мы провели кросс-валидацию на получившихся подвыборках.

Эксперименты в рамках RQ3

Для того чтобы ответить на RQ3, мы подготовили четыре группы морфемных разборов. В первую группу мы включили 50 случайных слов из набора данных Morphodict-T, во вторую — 50 случайных слов из набора данных Morphodict-K. Эта пара датасетов была выбрана для анализа, так как морфемные разборы между этими наборами данных различаются сильнее. В третью и четвертую группы мы включили по 50 слов из Morphodict-T и Morphodict-K, но не случайно отобранных, а таких, для которых обученный на соответствующем наборе данных ансамбль свёрточных нейронных сетей с определением типа морфемы сгенерировал некорректный разбор. Далее четырем экспертам было предложено сделать морфемный разбор для каждого из этих 200 слов в соответствии с парадигмой разметки: слова из групп 1 и 3 должны были быть разобраны согласно логике «Словообразовательного словаря русского языка», а слова из групп 2 и 4 — в соответствии со «Словарем морфем русского языка». Эксперты были заранее ознакомлены с использовавшимися при создании этих словарей принципами разметки, однако при этом они не могли использовать внешние источники информации (например, сами словари) в ходе разметки. Для достижения более объективных результатов слова из одного набора данных были перемешаны: группы 1 и 3 объединены в одну, группы 2 и 4 — в другую. После завершения разметки мы разделили слова на исходные группы и вычислили для каждой из групп степень согласованности экспертов между собой, а также оценили соответствие экспертной разметки словарной.

Метрики качества

Для оценки качества автоматической разметки в случае алгоритмов с определением типа морфем мы использовали метрики, предложенные в работе Sorokin & Kravtsova (2018): точность (Precision), полнота (Recall) и F-мера для границ морфем, посимвольная точность (Accuracy) и доля полностью верных разборов (WordAccuracy). Для оценки качества алгоритмов без определения типов мы вычисляли посимвольную точность (Accuracy), предварительно конвертировав разметку в BMES-формат, и долю полностью верных разборов (WordAccuracy). Для алгоритма DeepSPIN-3 мы дополнительно вычисляли Invalid — долю разборов, не совпадающих при конкатенации с исходным словом (для остальных алгоритмов в силу их природы эта метрика лишена смысла).

РЕЗУЛЬТАТЫ

Эксперименты в рамках RQ1

Результаты оценки моделей LSTM и GBDT представлены в Таблице 3 (для каждой пары набор данных + модель наибольшее значение соответствующей метрики выделено **жирным** шрифтом). Можно отметить, что для алгоритма LSTM использование дополнительной информации привело к снижению качества. Для алгоритма GBDT качество разметки выросло в двух из трех случаев, однако оно было незначительным. Кроме того, качество моделей GBDT значительно ниже, чем у аналогичных моделей LSTM. В связи с этим было решено учитывать для алгоритмов GBDT и LSTM только результаты, полученные без использования дополнительной информации при обучении.

Результаты для всех шести исследованных алгоритмов приведены в Таблицах 4 (алгоритмы с определением типа морфем) и 5 (алгоритмы без определения типа морфем; TCNN — модели TorchCNN, Mbert — модели Mbert, DS-3 — модели

[6] MorphemeBert (Библиотека Python). A. Sorokin. https://github.com/AlexeySorokin/MorphemeBert

Таблица 3

Сравнение качества автоматической разметки при помощи моделей LSTM и GBDT, обученных с/без дополнительной информации

Метрика	Вариант	LSTM			GBDT		
		Morphodict K	Morphodict T	CrossLexica	Morphodict K	Morphodict T	CrossLexica
Accuracy	Только разбор	**96.61**	**95.56**	**96.88**	88.84	86.88	92.26
	Лемма и часть речи	96.00	95.41	96.54	**88.96**	**87.29**	**92.37**
	Вся информация	96.07	95.40	96.22	88.93	86.91	92.10
WordAccuracy	Только разбор	**88.02**	**84.25**	**89.82**	64.43	58.63	75.25
	Лемма и часть речи	86.13	83.78	88.99	64.79	**60.01**	**75.62**
	Вся информация	86.42	83.75	87.97	**64.84**	59.14	75.06

Таблица 4

Сравнение качества разметки алгоритмов с определением типа морфем при случайном разбиении набора данных на подвыборки в ходе кросс-валидации

Метрика	Morphodict-K			Morphodict-T			CrossLexica		
	CNN	LSTM	GBDT	CNN	LSTM	GBDT	CNN	LSTM	GBDT
Precision	**98.58**	98.00	91.88	**97.79**	97.22	89.62	**98.74**	98.03	93.50
Recall	**98.74**	98.30	94.69	**98.38**	97.54	93.34	**99.04**	98.33	96.85
F-measure	**98.66**	98.15	93.26	**98.09**	97.38	91.44	**98.89**	98.18	95.14
Accuracy	**97.40**	96.61	88.84	**96.61**	95.56	86.88	**98.10**	96.88	92.26
WordAccuracy	**90.82**	88.02	64.43	**88.49**	84.25	58.63	**93.60**	89.82	75.25

Таблица 5

Сравнение качества разметки алгоритмов без определения типа морфем при случайном разбиении набора данных на подвыборки в ходе кросс-валидации

Метрика	Morphodict-K			Morphodict-T			CrossLexica		
	TCNN	MBert	DS-3	TCNN	MBert	DS-3	TCNN	MBert	DS-3
Invalid	-	-	12.22	-	-	11.32	-	-	17.02
Accuracy	97.43	**97.65**	86.07	96.80	**97.04**	86.21	98.01	**98.14**	81.83
WordAccuracy	89.42	**90.34**	80.89	86.00	**87.16**	78.28	91.99	**92.52**	78.43

DeepSPIN-3). Эксперимент показал, что среди алгоритмов с определением типа морфем однозначным лидером является ансамбль CNN. В случае алгоритмов без определения типа морфем ансамбли свёрточных нейронных сетей показали схожие результаты с небольшим преимуществом MorphemeBERT. Алгоритм DeepSPIN-3 в 11–17 % случаев сгенерировал разборы, не совпадающие с исходным словом, а качество разметки с его помощью оказалось ниже на 9–14 %.

Эксперименты в рамках RQ2

Результаты экспериментов с разбиением наборов данных по корням представлены в Таблицах 6 (алгоритмы с определением типа морфем) и 7 (алгоритмы без определения типа морфем). В дополнительной строке WA Drop приведена информация о снижении качества разметки в терминах WordAccuracy по отношению к эксперименту со случайным разбиением (в процентах, при этом качество разметки в случае случайного разбиения принято за 100 %). Из данных видно, что в случае свёрточных ансамблей и сетей LSTM снижение составляет 20–23 %, в случае алгоритма GBDT — 9–20 %, тогда как DeepSPIN-3 демонстрирует резкий рост доли разборов, не совпадающих с исходным словом. Сравнивая алгоритмы TorchCNN и MorphemeBERT, можно заметить, что в двух случаях из трех падение качества меньше у MorphemeBERT, причем разница становится более заметной с уменьшением размеров обучающей выборки.

Таблица 6

Сравнение качества разметки алгоритмов с определением типа морфем при разбиении набора данных на подвыборки по корням в ходе кросс-валидации

Метрика	Morphodict-K			Morphodict-T			CrossLexica		
	CNN	LSTM	GBDT	CNN	LSTM	GBDT	CNN	LSTM	GBDT
Precision	95.35	93.91	90.79	94.46	93.89	88.61	94.67	93.95	90.25
Recall	95.04	94.32	92.61	94.96	93.21	92.09	95.68	94.09	93.33
F-measure	95.19	94.11	91.69	94.71	93.54	90.32	95.17	94.02	91.77
Accuracy	91.30	89.64	86.58	90.16	88.41	84.87	91.28	89.53	87.01
WordAccuracy	72.63	67.80	58.67	70.53	65.47	53.72	74.14	69.48	60.08
WA Drop	20.03%	22.97%	8.95%	20.30%	22.30%	8.37%	20.79%	22.64%	20.16%

Таблица 7

Сравнение качества разметки алгоритмов без определения типа морфем в случае разбиения набора данных на подвыборки по корням в ходе кросс-валидации

Метрика	Morphodict-K			Morphodict-T			CrossLexica		
	TCNN	MBert	DS-3	TCNN	MBert	DS-3	TCNN	MBert	DS-3
Invalid	-	-	74.52	-	-	56.06	-	-	84.49
Accuracy	92.03	92.37	22.41	91.99	91.90	39.09	92.45	93.32	13.73
WordAccuracy	69.69	71.03	14.55	67.63	67.24	25.59	71.03	74.03	9.20
WA Drop	22.06%	21.37%	73.96%	21.36%	22.85%	54.65%	22.79%	19.98%	83.22%

Эксперименты в рамках RQ3

В Таблицах 8–11 представлены результаты сравнения экспертной и эталонной разметок для групп 1–4 соответственно. В каждой ячейке таблицы приведены значения Accuracy и WordAccuracy, разделенные символом «|». Интересными нам показались следующие наблюдения:

(1.) Качество экспертной разметки сопоставимо с качеством разметки при помощи ансамбля CNN.
(2.) Для всех четырех групп можно установить единый порядок экспертов по убыванию качества разметки: Эксперт 3 > Эксперт 4 > Эксперт 2 > Эксперт 1.
(3.) Согласованность экспертов между собой часто ниже, чем с эталонной разметкой, что может свидетельствовать о том, что характер различий с эталонной разметкой зависит от конкретного эксперта.
(4.) Согласованность экспертов между собой и с эталонной разметкой гораздо больше зависит от принципа выбора слов для группы, чем от источника слов: для случайно выбранных слов качество экспертной разметки и согласованность экспертов между собой заметно выше. Более того, как и в случае автоматической разметки, качество экспертной разметки немного выше для групп, опирающихся на Morphodict-K.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Эксперименты в рамках RQ1

Так как лучшие результаты среди обеих групп алгоритмов были достигнуты при помощи ансамблей свёрточных нейронных сетей, мы провели дополнительное исследование ошибок, допущенных CNN-моделями.

Хотя задача сегментации с определением типов морфем труднее задачи сегментации без определения, лучших результатов среди всех шести алгоритмов, согласно метрикам Accuracy и WordAccuracy, удалось добиться именно при помощи модели CNN. Модели TorchCNN и MorphemeBERT уступили в качестве разметки, несмотря на сходную архитектуру моделей. Мы связываем это с двумя факторами: во-первых, реализация алгоритма из оригинальной работы (Sorokin & Kravtsova, 2018) содержит ряд эвристик постобработки, направленных на исправление некоторых типов ошибок; во-вторых, для реализации были использованы разные фреймворки (TensorFlow[7] в первом случае, PyTorch[8] во втором) и разные версии программных пакетов.

Ранее в работе Sorokin & Kravtsova (2018) было установлено, что некоторые из ошибок алгоритма связаны не с не-

[7] TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. A. Martin et al. https://www.tensorflow.org/

[8] PyTorch. A. Paszke et al. https://pytorch.org/

Таблица 8

Значения Accuracy и WordAccuracy для экспертной разметки относительно эталонной разметки и относительно прочих экспертов. Группа 1: Morphodict-T, случайные слова

	Эталон	Эксперт 1	Эксперт 2	Эксперт 3	Эксперт 4
Эталон	-	90.79 | 70	90.79 | 72	97.05 | 92	96.69 | 88
Эксперт 1	90.79 | 70	-	89.87 | 66	89.69 | 68	92.82 | 72
Эксперт 2	90.79 | 72	89.87 | 66	-	89.69 | 70	93 | 76
Эксперт 3	97.05 | 92	89.69 | 68	89.69 | 70	-	94.66 | 84
Эксперт 4	96.69 | 88	92.82 | 72	93 | 76	94.66 | 84	-

Таблица 9

Значения Accuracy и WordAccuracy для экспертной разметки относительно эталонной разметки и относительно прочих экспертов. Группа 2: Morphodict-K, случайные слова

	Эталон	Эксперт 1	Эксперт 2	Эксперт 3	Эксперт 4
Эталон	-	78.18 | 36	83.64 | 44	95.35 | 86	92.12 | 74
Эксперт 1	78.18 | 36	-	83.84 | 52	78.99 | 38	82.42 | 44
Эксперт 2	83.64 | 44	83.84 | 52	-	85.05 | 52	84.65 | 52
Эксперт 3	95.35 | 86	78.99 | 38	85.05 | 52	-	88.89 | 68
Эксперт 4	92.12 | 74	82.42 | 44	84.65 | 52	88.89 | 68	-

Таблица 10

Значения Accuracy и WordAccuracy для экспертной разметки относительно эталонной разметки и относительно прочих экспертов. Группа 3: Morphodict-T, предположительно, сложные слова

	Эталон	Эксперт 1	Эксперт 2	Эксперт 3	Эксперт 4
Эталон	-	88.71 | 60	91.88 | 68	97.82 | 90	97.03 | 86
Эксперт 1	88.71 | 60	-	92.28 | 70	89.11 | 62	89.11 | 62
Эксперт 2	91.88 | 68	92.28 | 70	-	92.08 | 70	92.08 | 72
Эксперт 3	97.82 | 90	89.11 | 62	92.08 | 70	-	97.23 | 88
Эксперт 4	97.03 | 86	89.11 | 62	92.08 | 72	97.23 | 88	-

Таблица 11

Значения Accuracy и WordAccuracy для экспертной разметки относительно эталонной разметки и относительно прочих экспертов. Группа 4: Morphodict-K, предположительно, сложные слова

	Эталон	Эксперт 1	Эксперт 2	Эксперт 3	Эксперт 4
Эталон	-	82.05 | 46	82.69 | 44	95.51 | 86	94.02 | 80
Эксперт 1	82.05 | 46	-	76.71 | 32	80.77 | 46	85.04 | 54
Эксперт 2	82.69 | 44	76.71 | 32	-	81.62 | 42	83.97 | 50
Эксперт 3	95.51 | 86	80.77 | 46	81.62 | 42	-	88.68 | 62
Эксперт 4	94.02 | 80	85.04 | 54	83.97 | 50	88.68 | 62	-

достатками модели, а с непоследовательной или ошибочной разметкой в обучающих данных. Это подтверждается и нашими наблюдениями. Проанализировав ситуации, в которых автоматическая разметка отличалась от эталонной, мы обнаружили, что случаи, когда алгоритм правильно сегментировал слово, но неправильно выбрал типы морфем, достаточно редки и составляют около 9 % всех ошибочных разборов. Эти ошибки во многом объясняются именно непоследовательной разметкой в обучающих данных, в частности, путаницей между типами ROOT (корень) и PREF (приставка) в таких морфемах, как ультра-, мега-, супер- и подобных. В наборе данных Morphodict-K мы обнаружили семь случаев, когда *ультра-* была размечена как приставка, и два случая, когда эта же морфема была размечена как корень. Аналогичная ситуация наблюдается для *мега-*: шесть раз она классифицирована как приставка, четыре — как корень; для *супер-*: пять раз как приставка, десять раз как корень. Во всех этих случаях нам не удалось обнаружить причин, побудивших авторов Morphodict-K принять то или иное решение. Таким образом, следует считать, что, за исключением случаев непоследовательной разметки в обучающих данных, задача автоматического выбора типов

морфем для сегментированных слов может быть решена практически без ошибок.

Необходимость повышения согласованности обучающей выборки подтверждается ошибками, связанными с различной дискретностью членения суффиксов. Около 20 % различий между сгенерированными разборами и эталонными связано с объединением пары суффиксов в один, например, *-н-ик* и *-ник*. Оба варианта представлены в Morphodict-K, например, вечер:*ROOT*/ник:*SUFF,* *о:PREF/город:ROOT/ник:SUFF,* *борт:ROOT/ник:SUFF,* но при:*PREF/кла:ROOT/д:SUFF/н:SUFF/ик:SUFF,* не:*PREF/год:ROOT/н:SUFF/ик:SUFF,* еже:*PREF/год:ROOT/н:SUFF/ик:SUFF*.

Как и в работе Sapin & Bolshakova (2019a), часть ошибок автоматической разметки может быть исправлена простыми эвристическими правилами, опирающимися на автоматически присвоенные морфологические признаки слов. Например, замена присвоенных типов морфем END на SUFF для неизменяемых частей повысила долю полностью верных разборов примерно на 0,2 %. В то же время использование морфологических признаков, по-видимому, не является перспективным способом повышения качества. Об этом можно судить по результатам сравнения моделей LSTM и GBDT, обученным с добавлением морфологической информации и без неё: значительного повышения качества удалось добиться только в случае алгоритма GBDT и набора данных Morphodict-T, тогда как в остальных случаях добавление информации либо слабо влияло на результат, либо вовсе приводило к снижению качества.

Значительное число ошибок, допущенных моделями, связано с отсутствием учета семантики слов и обработкой аббревиатур и сокращений, например, за:*PREF*/влаб:*ROOT* вместо эталонного зав:*ROOT*/лаб:*ROOT* (от <u>зав</u>едующий <u>лаб</u>ораторией), во:*ROOT*/ен:*SUFF*/к:*SUFF*/ом:*SUFF* вместо эталонного во:*ROOT*/ен:*SUFF*/ком:*ROOT* (от <u>воен</u>ный <u>ком</u>иссар). Интересно, что в некоторых подобных случаях автоматическая разметка теоретически возможна, например, пере:*PREF*/дом:*ROOT* как образованное от *дом:ROOT* при помощи приставки *пере-* аналогично пере:*PREF*/груз:*ROOT* от груз:*ROOT* (корректным разбором в данном случае является перед:ROOT/ом:SUFF); не:*PREF*/суш:*ROOT*/к:*SUFF*/а:*END* может быть образовано от суш:ROOT/к:SUFF/а:END аналогично не:PREF/у:PREF/вер:ROOT/енн:SUFF/ость:SUFF от у:PREF/вер:ROOT/енн:SUFF/ость:SUFF (корректным разбором в данном случае является нес:ROOT/ушк:SUFF/а:END).

Как и в работах Bolshakova & Sapin (2019a), Bolshakova & Sapin (2019b), большинство ошибок автоматической разметки связано с определением границ корневых морфем. Следует предположить, что прогресса в этой области можно добиться с использованием моделей семантических векторов, предобученных на больших текстовых корпусах. Это предположение подкрепляется сравнением алгоритмов TorchCNN и MorphemeBERT. Обладая идентичной архитектурой модели, алгоритм MorphemeBERT позволил достичь прироста качества разметки на 0,5–1 % в терминах WordAccuracy, что согласуется с результатами, полученными в работе Sorokin (2021) на материале шести других языков.

Следует также отметить, что нам не удалось добиться качества разметки моделями LSTM и GBDT, сопоставимого с полученным в оригинальных работах (Bolshakova & Sapin, 2019a; Bolshakova & Sapin, 2019b). В нашем случае сети LSTM не превзошли ансамбль свёрточных нейронных сетей ни для какого набора данных (впрочем, следует упомянуть, что мы не проводили экспериментов с ансамблированием моделей LSTM). Еще одним отличием стало слабое влияние морфологической информации на качество разметки. Мы полагаем, что основной причиной этих расхождений, как и в случае сравнения ансамблей свёрточных нейронных сетей, стали различия в версиях использованных библиотек. В то же время как и в работах Bolshakova & Sapin (2019a), Bolshakova & Sapin (2019b), качество автоматической сегментации на наборе данных CrossLexica оказалось выше, чем на наборе данных, основанном на «Словообразовательном словаре русского языка». Таким образом, несмотря на некоторые различия, наши результаты достаточно хорошо согласуются с ранее полученными, обобщая их на большее количество алгоритмов и наборов данных.

Качество разметки алгоритмом DeepSPIN-3 также оказалось значительно ниже в сравнении с оригинальной работой. Это в первую очередь связано с существенными различиями в принципах построения наборов данных: на соревновании SIGMORPHON набор данных для русского языка был примерно в 10 раз больше, чем Morphodict-K, но значительный процент составляли не леммы, а словоформы, причем разные формы одного и того же слова могли присутствовать как в обучающей, так и в тестовой выборках. Выбор такого подхода к построению набора данных может оказаться эффективным для использования моделей в качестве токенизаторов, но не до конца ясно, насколько он применим для автоматического расширения морфемных словарей. В будущем мы планируем провести дополнительные исследования в этом направлении, расширяя наш набор данных за счет автоматически собранных и размеченных словоформ.

Эксперименты в рамках RQ2

Анализ качества генерируемой разметки при разделении на обучающую и тестовую выборки по корням показал, что качество всех рассмотренных алгоритмов значительно снижается при такой постановке задачи, что является критичным для автоматического расширения морфемных словарей. Это согласуется с выводами, полученными в работе Garipov et al. (2023) для ансамбля свёрточных нейронных сетей, и расширяет исследование, охватывая несколько алгоритмов, которые ранее не исследовались с этой точки зрения. Ошибки, допущенные ансамблем CNN при данной

постановке задачи, отличаются от ошибок, возникающих при случайном разбиении набора данных: в ряде случаев модель предпринимает попытку вычленить морфемы, уже встречавшиеся в обучающей выборке. Так, модель генерирует разбор при:*PREF*/бран:*ROOT*/н:*SUFF*/ый:*END*, опираясь на имеющийся в обучающей выборке корень *-бран-* (ср. не:*PREF*/воз:*PREF*/бран:*ROOT*/н:*SUFF*/ый:*END*), при эталонном разборе при:*PREF*/бр:*ROOT*/а:*SUFF*/нн:*SUFF*/ый:*END*. Как и в случае сегментации сокращений, повышения качества разметки в подобных случаях следует ожидать от применения предобученных языковых моделей, особенно при работе с малыми размерами обучающих выборок.

Эксперименты в рамках RQ3

Насколько нам известно, сравнительный анализ экспертной и автоматической разметки в рамках задачи построения морфемных разборов на материале русского языка ранее не проводился. Поэтому мы решили провести подробный анализ ошибок, допущенных экспертами. Мы обнаружили, что в большинстве случаев эталонный разбор был подтвержден хотя бы одним экспертом. При этом как минимум два эксперта построили разбор, совпадающий с эталонным, в 50 (из 50) случаях для группы 1, в 40 случаях для группы 2, в 45 случаях для группы 3 и в 36 случаях для группы 4. Только для шести слов эталонный разбор не был построен ни одним из экспертов: усердн:ROOT/ый:END, чет:ROOT/в:SUFF/ер:SUFF/ич:SUFF/н:SUFF/ый:END (группа 2), о:PREF/свежева:ROOT/нн:SUFF/ый:END, чет:ROOT/в:SUFF/ер:SUFF/ик:SUFF (группа 3), короб:ROOT/чат:SUFF/ый:END, не:PREF/про:PREF/долж:ROOT/и:SUFF/тельн:SUFF/ый:END (группа 4). Следует отметить, что некоторые из этих шести случаев могут быть связаны с ошибками в эталонной разметке: избыточной дробностью при выделении корня в случаях чет:ROOT/в:SUFF/ер:SUFF/ич:SUFF/н:SUFF/ый:END и чет:ROOT/в:SUFF/ер:SUFF/ик:SUFF, недостаточной дробностью в случае усердн:ROOT/ый:END (ср. усердие без суффикса *-н-*), слиянием суффиксов в случаях короб:ROOT/чат:SUFF/ый:END и не:PREF/про:PREF/долж:ROOT/и:SUFF/тельн:SUFF/ый:END (ср. с имеющимися в Morphodict-K сум:ROOT/ч:SUFF/ат:SUFF/ый:END и у:PREF/по:PREF/доб:ROOT/и:SUFF/тель:SUFF/н:SUFF/ый:END).

Мы распределили различия между экспертной и эталонной разметкой по нескольким группам (в скобках приведено число подобных расхождений):

- Группа 1 (Morphodict-T, случайные слова): root vs root+suff (9), root vs pref+root (8), **дробность корня** (6), suff vs suff+suff (5)
- Группа 2 (Morphodict-K, случайные слова): suff vs suff+suff (21), root vs root+suff (13), root vs pref (4)
- Группа 3 (Morphodict-T, «сложные» случаи): root vs root+suff (29), **дробность корня** (14), root vs pref+root (11), suff vs suff+suff (10)
- Группа 4 (Morphodict-K, «сложные» случаи): root vs root+suff (23), suff vs suff+suff (12), root vs pref+root (8), root vs root+link (7)

Здесь **root vs root+suff** относится к случаям, когда разборы различаются дополнительным суффиксом, вычлененным из корня, аналогично в случаях **root vs pref+root** (вычленяется приставка) и **root vs root+link** (вычленяется соединительная гласная). В случае **suff vs suff+suff** один суффикс разбивается на два, а в случае **root vs pref** разборы различаются по типу выделенной морфемы: корнем или приставкой. Наконец, в случаях, обозначенных как **дробность корня**, длинный корень разбивается на несколько (более двух) морфем. Результаты проведенного анализа подтверждают нашу гипотезу: степень дробности разборов в различных источниках слабо формализована и приводит к несогласованности разметки. Наиболее распространенные случаи расхождения как между экспертами, так и при сравнении экспертной разметки с эталонной, а именно -тель:*SUFF*/н:*SUFF*- и -тельн:*SUFF*-, -н:*SUFF*/ик:*SUFF*- и -ник:*SUFF*-, -ич:*SUFF*/а:*SUFF*- и -ича:*SUFF*-, встречаются в наборах данных в обоих вариантах сегментации.

К сожалению, число слов, отмеченных экспертами как неизвестные (содержащие неизвестные корни), оказалось незначительным, из-за чего нам не удалось сравнить качество экспертной разметки с автоматической. В дальнейшем мы планируем провести дополнительно исследовать это направление.

Ограничения исследования

Основным ограничением исследования является использование словарей, содержащих исключительно или почти исключительно леммы, а не словоформы. Это связано с тем, что нам не удалось найти морфемные словари словоформ достаточного объема для обучения моделей. Однако в прикладных задачах зачастую требуется анализировать именно словоформы. В связи с этим представляется необходимым поиск или создание морфемного словаря словоформ и повторная оценка алгоритмов на его материале.

Кроме того, нам не удалось сравнить работу алгоритмов и экспертов на словах с незнакомыми корнями, так как в используемых словарях оказалось недостаточно слов с корнями, незнакомыми экспертам.

ЗАКЛЮЧЕНИЕ

Автоматическое построение морфемных разборов востребовано для задач изучения языка и при построении инструментов автоматической обработки естественного языка. В последние десятилетия было предложено множество алгоритмов для решения этих задач. Однако сравнение качества подходов затруднено из-за различий в использованных наборах данных и формулировках задачи. В нашем исследо-

вании мы провели подробное сравнение шести актуальных алгоритмов морфемной сегментации на материале русского языка с использованием трех морфемных словарей, различающихся принципами морфемного членения. Это позволило нам получить репрезентативные результаты и определить, как соотносится качество этих алгоритмов в различных условиях. Мы сравнили качество автоматической и экспертной разметки, чтобы определить перспективные направления улучшения алгоритмов. Кроме того, мы исследовали ранее выявленный недостаток моделей: значительное снижение качества разметки при обработке слов с корнями, отсутствующими в обучающей выборке.

Мы обнаружили, что лучшего качества разметки на материале всех наборов данных удается добиться при помощи ансамбля свёрточных нейронных сетей, и это качество может быть повышено за счет использования векторных представлений слов, полученных из BERT-подобных моделей. Анализ ошибок алгоритма показал, что многие из них связаны с внутренней несогласованностью обучающих данных, отсутствием учета семантики при сегментации, обработкой сокращений и аббревиатур. Результаты нашего эксперимента подтвердили снижение качества разметки всех исследованных алгоритмов при обработке незнакомых корней, что затрудняет использование этих методов для автоматического пополнения существующих морфемных словарей.

Сравнение автоматической и экспертной разметки показало, что при обработке случайных слов алгоритмы уже достигли экспертного уровня согласно автоматически вычисляемым метрикам. Ошибки экспертов обычно связаны с принятием локальных решений о степени дробности сегментации, что, на наш взгляд, иллюстрирует тот факт, что сегментация морфем для русского языка часто основана на прецедентах, то есть опирается на ранее размеченные случаи, и не может быть построена исключительно на основании заявленной парадигмы членения.

Таким образом, дальнейшие исследования в этой области должны быть направлены не на повышение среднего качества разметки, но на решение основных обнаруженных недостатков автоматической разметки: плохого распознавания незнакомых корней, сокращений и аббревиатур. Вероятно, учет семантики и распознавание аббревиатур может быть реализован при помощи языковых моделей, предобученных на больших корпусах текстов. Кроме того, будущие исследования должны быть проведены не только на материале словарей лемм, но и на словоформных словарях. На сегодняшний день такие исследования затруднены из-за недостатка размеченных данных, однако последние работы открывают перспективы для развития в этом направлении.

БЛАГОДАРНОСТИ

Мы выражаем свою признательность Дмитрию Сичинаве за рекомендации при проведении исследований и Софье Чижевской за помощь в подготовке англоязычной версии этого текста.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

ВКЛАД АВТОРОВ

Морозов Дмитрий Алексеевич: концептуализация, методология, программное обеспечение, исследование, написание - подготовка оригинального проекта.

Гарипов Тимур Александрович: методология, программное обеспечение.

Ляшевская Ольга Николаевна: обработка данных, исследование, написание - рецензирование и редактирование.

Савчук Светлана Олеговна: обработка данных.

Иомдин Борис Леонидович: сбор данных, написание - рецензирование и редактирование.

Глазкова Анна Валерьевна: методология, проверка достоверности, написание - рецензирование и редактирование.

ЛИТЕРАТУРА

Bakulina, G. A. (2012). Morfemnyy razbor slova: novye podkhody — novye vozmozhnosti [Morpheme segmentation: new approaches - new opportunities]. *Nachal'naya shkola*, (4), 29–32.

Batsuren, K., Bella, G., Arora, A., Martinovic, V., Gorman, K., Žabokrtský, Z., Ganbold, A., Dohnalová, Š., Ševčíková, M., Pelegrinová, K., Giunchiglia, F., Cotterell, R., & Vylomova, E. (2022). The SIGMORPHON 2022 shared task on morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 103–116). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.sigmorphon-1.11

Bodnár, J. (2022). JB132 submission to the SIGMORPHON 2022 shared task 3 on morphological segmentation. *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 152–156). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.sigmorphon-1.17

Bolshakov, I.A. (2013). Krossleksika: Universum sviazi mezhdu russkimi slovami [Crosslexica: a universe of links between russian words]. *Biznes-informatika*, *3*(25), 12–19.

Bolshakova, E., Sapin, A. (2019). Bi-LSTM model for morpheme segmentation of russian words. In Ustalov, D., Filchenkov, A., Pivovarova, L. (Eds.), *Artificial Intelligence and Natural Language. AINL 2019. Communications in Computer and Information Science* (pp. 151-160). Springer. https://doi.org/10.1007/978-3-030-34518-1_11

Bolshakova, E., Sapin, A. (2021). Building a Combined morphological model for Russian word forms. In Burnaev, E. et al. (Eds), *Analysis of Images, Social Networks and Texts. AIST 2021. Lecture Notes in Computer Science* (vol. 13217, pp. 45-55). Springer. https://doi.org/10.1007/978-3-031-16500-9_5

Bolshakova, E.I., & Sapin, A.S. (2019). Comparing models of morpheme analysis for Russian words based on machine learning. *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference Dialogue 2019* (pp. 104-113). Russian State University for the Humanities.

Creutz, M., & Lagus, K. (2002). Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning* (pp. 21–30). Association for Computational Linguistics. https://doi.org/10.3115/1118647.1118650

Cotterell, R., Vieira, T., & Schütze, H. (2016). A joint model of orthography and morphological segmentation. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 664–669). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1080

Garipov, T., Morozov, D., & Glazkova, A. (2023). Generalization ability of CNN-based morpheme segmentation. *2023 Ivannikov Ispras Open Conference (ISPRAS)* (pp. 58–62). IEEE https://doi.org/10.1109/ISPRAS60948.2023.10508171

Girrbach, L. (2022). SIGMORPHON 2022 shared task on morpheme segmentation submission description: Sequence labelling for word-level morpheme segmentation. *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 124–130). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.sigmorphon-1.13

Grönroos, S.-A., Virpioja, S., & Kurimo, M. (2020). Morfessor EM+Prune: Improved subword segmentation with expectation maximization and pruning. *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 3944–3953). European Language Resources Association.

Imani, A., Lin, P., Kargaran, A. H., Severini, S., Sabet, M. J., Kassner, N., Ma, C., Schmid, H., Martins, A., Yvon, F., & Schütze, H. (2023). Glot500: Scaling multilingual corpora and language models to 500 languages. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (vol. 1: Long Papers, pp. 1082–1117). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.61

Iomdin, B. L. (2019). How to define words with the same root? *Russian Speech*, (1), 109–115. https://doi.org/10.31857/S013161170003980-7

Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (vol. 1: Long Papers, pp. 66–75). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1007

Kuratov, Y. & Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for Russian language. *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference Dialogue 2019* (pp. 333–339). Russian State University for the Humanities.

Kuznetsova, A. I. & Efremova, T. F. (1986). *Dictionary of morphemes of the Russian language*. Russkii yazyk.

Levine, L. (2022). Sharing data by language family: Data augmentation for romance language morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 117–123). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.sigmorphon-1.12

Matthews, A., Neubig, G., & Dyer, C. (2018). 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* (vol. 1, pp. 1435–1445). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1130

Morozov, D. A., Smal, I. A., Garipov, T. A., & Glazkova, A. V. (2024). Keywords, morpheme parsing and syntactic trees: Features for text complexity assessment. *Modeling and Analysis of Information Systems*, *31*(2), 206–220. https://doi.org/10.18255/1818-1015-2024-2-206-220

Peters, B. & Martins, A. F. T. (2022). Beyond characters: Subword-level morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 131–138). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.sigmorphon-1.14

Pranjić, M., Robnik-Šikonja M., & Pollak, S. (2024). LLMSegm: Surface-level morphological segmentation using large language model. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (pp. 10665–10674). ELRA and ICCL.

Savchuk, S. O., Arkhangelskiy, T., Bonch-Osmolovskaya, A. A., Donina, O. V., Kuznetsova, Yu. N., Lyashevskaya, O. N., Orekhov, B. V., & Podryadchikova, M. V. (2024). Russian national corpus 2.0: New opportunities and development prospects. *Voprosy Jazykoznanija*, *2,* 7–34. https://doi.org/10.31857/0373-658X.2024.2.7-34

Schuster, M. & Nakajima, K. (2012). Japanese and Korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing* (pp. 5149–5152). IEEE. https://doi.org/10.1109/ICASSP.2012.6289079

Sorokin, A. & Kravtsova, A. (2018). Deep convolutional networks for supervised morpheme segmentation of Russian language. In D. Ustalov, A. Filchenkov, L. Pivovarova, & J. Žižka, (Eds.), *Artificial Intelligence and Natural Language* (pp. 3-10). Springer. https://doi.org/10.1007/978-3-030-01204-5_1

Sorokin, A. (2022). Improving morpheme segmentation using BERT embeddings. In E. Burnaev, D. Ignatov, S. Ivanov, M. Khachay, O. Koltsova, A. Kutuzov, S.Kuznetsov, N. Loukachevitch, A. Napoli, A. Panchenko, P. Pardalos, J. Saramäki, A. Savchenko, E. Tsymbalov, & E. Tutubalina, (Eds.), *Analysis of images, social networks and texts* (pp. 148-161). Springer. https://doi.org/10.1007/978-3-031-16500-9_13

Tikhonov, A. N. (1990). *Slovoobrazovatel'nyi slovar' russkogo yazyka* [Word Formation Dictionary of Russian language]. Russkiy yazyk.

Vinokur, G. O. (1946). *Zametki po russkomu slovoobrazovaniyu* [Notes on Russian word formation]. *Izvestiya Akademii nauk SSSR. Seriya literatury i yazyka*, *V*(4), 317-317.

Wehrli, S., Clematide, S., & Makarov, P. (2022). 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* (pp. 212–219). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.sigmorphon-1.21

Zundi, T. & Avaajargal, C. (2022). Word-level Morpheme segmentation using Transformer neural network. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 139–143). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.sigmorphon-1.15