УДК 81'32+81'33

К ВОПРОСУ О ПРЕДСТАВЛЕНИИ СИНТАГМАТИЧЕСКИХ ОТНОШЕНИЙ МОРФЕМ В ВЕКТОРНЫХ ЯЗЫКОВЫХ МОДЕЛЯХ

Д. К. Родионова[1] [0009-0004-6296-8532], О. А. Митрофанова[2] [0000-0002-3008-5514]

[1, 2]*Санкт-Петербургский государственный университет,*
г. Санкт-Петербург, Россия

[1]*НИИ Исследовательская лаборатория им. П. Л. Чебышева,*
г. Санкт-Петербург, Россия

[1]rodionowadarja@yandex.ru, [2]o.mitrofanova@spbu.ru

Аннотация

В работе рассмотрено представление семантической структуры производных слов в языковых моделях, учитывающее внутрисловные синтагматические отношения между словообразовательными морфемами. Эксперименты проводились с привлечением морфемных моделей НейроКРЯ, а также моделей fastText и ruRoBERTa. Проверена гипотеза о композициональности производных слов, представляемых в виде агрегированных векторов морфем, а также выполнено сравнение представлений семантических отношений с помощью морфемных векторов fastText и стандартных векторов подслов в модели ruRoBERTa. Полученные результаты указывают на умеренную чувствительность векторов fastText к синтагматическим связям между морфемами и словообразовательным типам. Установлено также что агрегация морфемных векторов в fastText улучшает регистрацию семантических отношений между словами, связанными словообразовательными отношениями, по сравнению с агрегацией векторов подслов в модели ruRoBERTa.

Стандартные токенизаторы BPE (Byte-Pair Encoding) и WordPiece, применяемые в моделях семейства Transformer, являются слабоинтерпретируемыми в отношении языковых данных, поскольку в них сегменты слов не всегда соответствуют морфемам. Исследовательская проблема состоит в необходимости оценки того, в какой мере современные языковые модели способны регистрировать лингвистические признаки, характеризующие отношения производных слов в словообразовательных гнездах.

В работе оценена способность предсказывающих моделей распределенных векторных вложений воспроизводить синтагматические связи между морфемами внутри производных слов и на уровне словообразовательных гнезд в русском языке.

Полученные результаты стимулируют разработку нейросетевых архитектур, учитывающих синтагматические отношения между морфемами, совершенствование морфемных токенизаторов и их интеграцию в языковые модели.

***Ключевые слова**: языковая модель, морфемный анализ, словообразовательные способы, композициональность.*

ВВЕДЕНИЕ

На сегодняшний день ни одна задача обработки естественного языка не обходится без применения методов векторизации текстовых данных и интеграции больших языковых моделей в лингвистические процессоры. Современные подходы к анализу текстов пользуются большой популярностью благодаря появлению вычислительных ресурсов, позволяющих обработать большие объемы данных, что обеспечивает высокое качество моделей и верификацию результатов. В то же время все больше вопросов возникает в связи с интерпретируемостью внутренних представлений моделей и их соответствием языковым единицам различных уровней, в том числе морфем [1]. Токенизаторы классов BPE и WordPiece, используемые в моделях семейства Transformer, являются слабоинтерпретируемыми, поскольку выделяют сегменты слов, не всегда соответствующие морфемам. Во многих работах было показано положительное влияние морфемной токенизации на качество генерации текстов с использованием приемов перефразирования, суммаризации, упрощения в языках с богатыми словообразованием и словоизменением (русский, белорусский, сербский, чешский, финский, эстонский и т. д.). Кроме того, морфемный анализ может положительно влиять на качество морфологической аннотации текстов и генерации морфологических форм слов [2–6]. Несмотря на это, исследование внутренней структуры слова в русскоязычных языковых моделях недостаточно широко представлено в публикациях.

Цель настоящего исследования состояла в оценке способности предсказывающих моделей распределенных векторных вложений воспроизводить синтагматические связи между морфемами внутри производных слов и на уровне словообразовательных гнезд в русском языке. В ходе исследования проверялась гипотеза о композициональности производных слов при агрегации морфемных векторов.

В статье дан обзор аналогичных исследований, описан исследовательский набор данных, обоснован выбор моделей fastText и ruRoBERTa, представлены способы агрегации векторов производных слов, а также проведены анализ результатов сравнения агрегированных векторов для исследовательского набора данных в моделях и оценка способности моделей fastText и ruRoBERTa воспроизводить семантические отношения внутри словообразовательных гнезд. Полученные результаты подтверждают перспективность учета границ морфем в разработке токенизаторов для языковых моделей.

БЛИЗКИЕ ПО ТЕМАТИКЕ ИССЛЕДОВАНИЯ

Вычислительные аспекты нашего исследования согласуются с тенденциями развития языкового моделирования как задачи искусственного интеллекта, в то время как лингвистические основания связаны с особым направлением в формальной лингвистике, а именно с генеративной морфологией, применяющей аппарат формальных грамматик в описании процессов деривации [7], и теорией гипосинтаксиса, объясняющей природу синтагматических отношений между морфемами [8]. Для русского языка, для которого характерны богатая морфологическая система, развитые словоизменение и словообразование, особо важно, что производные слова обладают дискретной структурой как в плане выражения, так и в плане содержания. Значение производного слова возникает в результате воздействия словообразовательного аффикса на производящую основу. Это дает основания считать внутрисловные связи между морфемами разновидностью синтаксических отношений, поэтому, производное слово может рассматриваться как аналог словосочетания и предложения [9–12]. Интеграция генеративного и традиционного подходов к описанию семантики производного слова реализована в деривационных моделях, использующих падежно-ролевой подход [13–15]. Тем самым в указанных работах рассмотрена

проблема композициональности семантики производных слов, но вне задачи обучения и применения языковых моделей.

С возможностью введения в корпусы текстов словообразовательной разметки (в частности, в НКРЯ) и учета деривационных связей в компьютерных тезаурусах типа WordNet задача моделирования словообразовательных отношений стала более реалистичной. В условиях ограниченных обучающих данных применимы обучение без учителя и статистические подходы, в частности, в инструменте Morfessor [16] реализован алгоритм вероятностной сегментации слов на морфемы, адаптируемый к различным языкам (финский, турецкий, эстонский, русский и т. д.).

При наличии обучающих данных высокие результаты обеспечиваются алгоритмами глубинного обучения. В частности, для русского языка существует группа нейросетевых моделей, обученных под задачу морфемной сегментации и классификации: CNN, LSTM, GBDT, BERT [2–6]. Нейросетевая классификация морфем состоит в присвоении части слова одной из специальных меток: префикса, корня, суффикса, окончания и т. д.

В работах [2, 3] представлены программный комплекс RussianMorphParsing [17] и набор данных RuMorphs-Lemmas, в [6] – инструмент и модели ruMorpheme [18], в серии публикаций [4, 5] и репозитории Neuromodels [19] – нейросетевые модели семейства BERT и словари, используемые в проекте НейроКРЯ. Недавно были представлены исследования, в которых рассматривались токенизаторы для моделей семейства Transformer, основанные на сегментации слов на морфемы [20–22]. Было показано, что благодаря такой стратегии они помогают повысить качество в решении различных лингвистических задач в отличие от обычных BPE-токенизаторов, которые при сегментации слов не учитывают границы морфем.

Несмотря на разнообразие решений задачи морфемной сегментации и классификации, до сих пор не решен вопрос о представлении синтагматических связей между морфемами в производных словах и отношений производности в словообразовательных гнездах. В настоящей работе предложено решение этих проблем.

ЭКСПЕРИМЕНТ

Данные

В качестве источника данных для серии экспериментов были использованы «Школьный словарь строения слов русского языка» З. А. Потихи объемом около 25 тыс. слов [23] и «Морфемно-орфографический словарь русского языка» А. Н. Тихонова объемом около 100 тыс. слов [24]. При отборе материала из этих источников учитывалась частотность целевых слов, а также репрезентативность их словообразовательных гнезд с точки зрения разнообразия словообразовательных способов. Мы также учитывали возможные разночтения в вариантах морфемной сегментации, представленных в разных источниках. По этим критериям были выбраны семь словообразовательных гнезд для существительных: *свет*, *лес*, *вода*, *дом*, *слово*, *земля* и *снег*. Объем гнезд для каждого производящего слова составлял примерно 50 лексических единиц. Общее число производных составляет более 350 лексических единиц. В каждом из гнезд представлены префиксально-суффиксальный, суффиксальный, префиксальный, сложно-суффиксальный словообразовательные способы, а также сложение основ (табл. 1), что позволило исследовать чувствительность языковых моделей к словообразовательным способам.

Табл. 1. Данные по словообразовательным гнездам.

Словообразовательный способ	СВЕТ	ЛЕС	ВОДА	ДОМ	СЛОВО	ЗЕМЛЯ	СНЕГ
Префиксно-суффиксальный	13	10	9	11	9	13	10
Суффиксальный	10	11	9	12	12	13	12
Префиксальный	6	5	1	0	1	2	0
Сложение основ	10	11	13	11	10	11	10
Сложно-суффиксальный	10	10	19	11	12	11	11
Общее количество	49	47	51	45	44	50	43

Умеренные объемы данных обусловлены тем, что на данном этапе исследования отсутствует такой инструмент, с помощью которого можно автоматизировать процесс сбора производных слов из предложенных выше словарей для составления гнезд. На величину гнезда влияет также исключение слов, которые

при одинаковом словообразовательном способе имеют различные окончания (например, в паре *светлый – светлая* оставляем первое слово).

Модели

Нейросетевые морфемные модели CNN, LSTM, GBDT, BERT в комбинации со словарными данными позволяют достичь при решении задачи морфемной сегментации значений F-меры на уровне 0.99. Модели MorphBERTa, разработанные НейроКРЯ [19], показывают на сегодняшний день наилучшие результаты в задачах определения морфемных границ и назначения морфемных меток. Однако, несмотря на свои преимущества, они имеют некоторые ограничения.

Во-первых, модели MorphBERTa не обучались для задачи распознавания границ предложений и не адаптированы для разрешения неоднозначности некоторых грамматических характеристик слов в контексте (например, словоформа *пора* в зависимости от синтаксической структуры предложения может быть аннотирована либо как предикативное наречие, либо как существительное).

Во-вторых, модели MorphBERTa не предназначены для распознавания словообразовательных способов (например, *учащийся* прич. → сущ.). Эти наблюдения требуют пересмотра исследовательского набора данных при подготовке экспериментов.

Для проверки гипотезы о композициональности производных слов при агрегации морфемных векторов мы рассмотрели группу моделей из семейства fastText [25], которые не были дообучены для обработки морфемной информации. Благодаря обучению на *n*-граммах (последовательностях графем внутри слов) модели fastText способны распознавать слова, отсутствующие в обучающих данных, и делать предсказания в отношении несловарных слов. Из предобученных моделей для русского языка были использованы *geowac_lemmas* и *geowac_tokens* с размером окна 5 и размерностью вектора 300 [26]. Дополнительно был проведен эксперимент с моделями Transformer для оценки способности моделей воспроизводить семантические отношения внутри словообразовательных гнезд. Из семейства BERT мы выбрали ruRoBERTa-large [27, 28] как альтернативу составной модели fastText и MorphBERTa, не содержащую информа-

цию о морфемном членении и разметке. В качестве токенов ruRoBERTa-large кодирует подслова. Токенизация проводится с помощью алгоритма BPE, который разбивает входные слова на подстроки и ранжирует их таким образом, что в словаре модели сохраняются наиболее частотные последовательности символов, которые далеко не всегда соответствуют морфемам.

Методы и метрики

В экспериментах были использованы следующие методы агрегации векторов производных слов. Для каждого слова в словообразовательном гнезде были сформированы три вектора: вектор производного слова, вектор из композиции морфем, а также вектор основы. Для вектора композиции агрегация проводилась одним из трех способов: это усреднение, сумма и выбор максимальной координаты. Далее вычисляли следующие косинусные метрики, которые затем собирались для каждого гнезда в отдельные выборки:

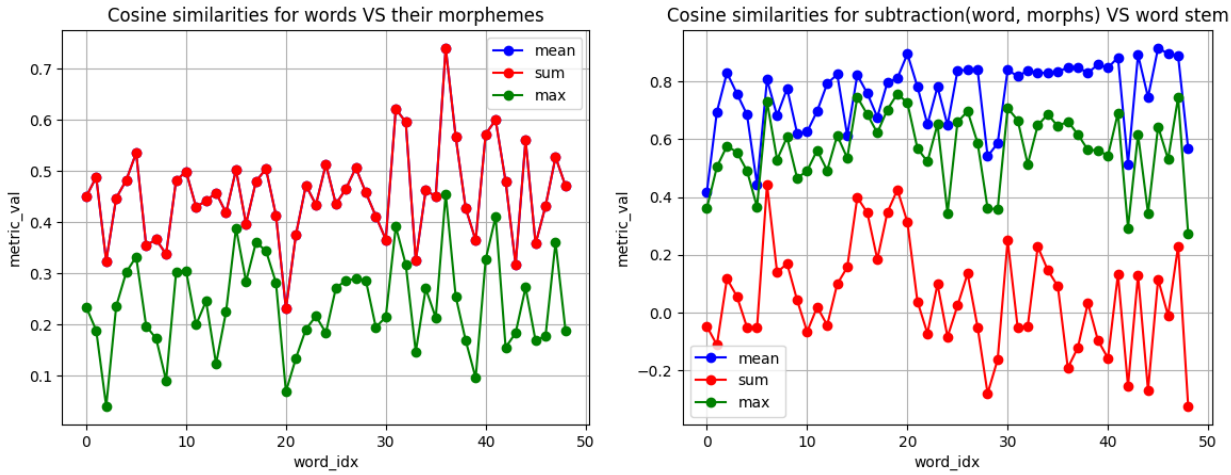
КМ-1: $\mathrm{cos}ine\big(w, aggr(m_i)\big)$;
КМ-2: $cosine(w\ -\ aggr(m_i), s)$, где
w – вектор слова, s – вектор основы слова,
$\{m_i\}$ – морфемный ряд, *aggr* = [*'mean'*, *'sum'*, *'max'*].

На первом этапе сравнивали способы агрегации векторов морфем в паре моделей fastText, из которых модель *geowac_tokens* была обучена на словоформах, а *geowac_lemmas* – на леммах. По данным, полученным по каждому из словообразовательных гнезд, выполняли дисперсионный анализ и его аналоги (тесты Краскела и Уелча) с целью проверки соотношения между словообразовательными способами и значениями косинусной метрики, а также выбора тех словообразовательных способов, которые лучше других представлены в моделях. Данный анализ проводился со значением *p*-value, равным 5%. Аналогичные шаги были также выполнены и для модели ruRoBERTa-large.

РЕЗУЛЬТАТЫ

Проверка гипотезы о композициональности производных слов при агрегации морфемных векторов

В ходе первого эксперимента было установлено, что обе модели fastText, обученные на словоформах и леммах, при использовании агрегации морфемных векторов методом усреднения дают наилучшие результаты. При этом значение косинусной метрики в целом не превышает 0.5, что означает умеренную степень близости между вектором слова и агрегированным вектором морфем. На рис. 1 представлены графики изменений значения косинусной метрики в словообразовательном гнезде слова *свет*.

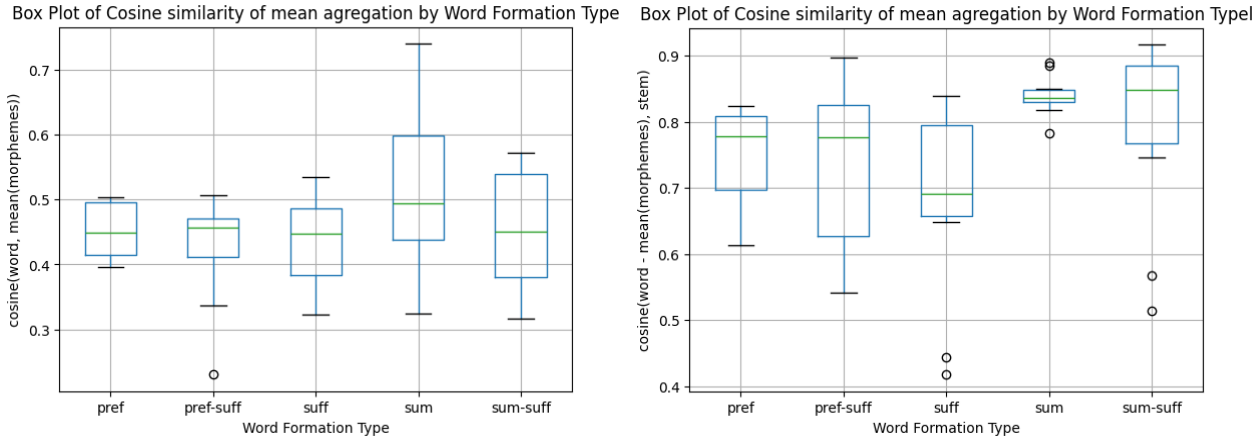


(а) (б)

Рис. 1. Косинусные значения для агрегаций векторов морфем в словообразовательном гнезде *свет*: а) КМ-1, б) КМ-2.

Было исследовано соотношение между словообразовательными способами и значениями косинусной метрики для агрегированных векторов. Отдельные словообразовательные способы и их значения метрики КМ-1 для словообразовательных гнезд представлены на ящиках с усами (рис. 2а – сравниваются векторы морфем, рис. 2б – сравниваются вектор основы и разность вектора слова и сводного вектора морфем). Следует заметить, что словообразовательный способ, связанный со сложением основ слов, показывает самые высокие результаты по метрике КМ-2 в случае агрегации морфемных векторов методом

усреднения. Это означает, что модели fastText могут обрабатывать многоосновные слова, имеющие слитное написание (например, *золотоискатель*, *Роспотребнадзор*). Однако такая закономерность не наблюдается для метрики КМ-1. Например, для слова *вода* наиболее высокие косинусные метрики соответствуют группе суффиксальной словообразовательной модели.



(а) (б)

Рис. 2. Распределение словообразовательных способов для словообразовательного гнезда *свет:* а) КМ-1, б) КМ-2.

Для проверки влияния способа словообразования на значение косинусной метрики может быть применен дисперсионный анализ при условии, что распределение исследуемых данных подчиняется нормальному закону. Если в выборках обнаруживались выбросы, часть из них исключалась, если они возникли вследствие ошибок модели НейроКРЯ. Например, для производящего слова *словарь* модель вернула морфемный ряд, состоящий только из одного корня *словарь*, что не соответствует правильному разбору, в котором выделяется суффикс *-арь*.

Для всех словообразовательных гнезд у fastText значения косинусных метрик КМ-2 оказались выше, чем КМ-1 (табл. 2). Более того, по результатам статистического сравнения двух моделей fastText для КМ-2 лучше всего подходит модель, обученная на леммах. Тем самым гипотеза о композициональности подтверждается при сравнении основы слова с разностью вектора слова и сводного вектора морфем. Однако, если исходить из значений косинусной метрики, связь

между вектором слова и агрегацией векторов морфем менее очевидна. Для определения сходства между агрегированными морфемами и словоизменительными аффиксами (прежде всего, флексии) рассчитывалась косинусная близость их векторов. Оказалось, что при усреднении векторов морфем выделялось только 30% флексий, для которых значения косинусной метрики были выше 0.7. Это указывает на слабую взаимосвязь между агрегированными морфемами и флексиями, а также позволяет предположить, что агрегированный вектор способен нести в себе более сложную информацию, чем вектор отдельной морфемы. Однако метрики модели ruRoBERTa-large имеют иные показатели: здесь значения КМ-1 выше, чем значения КМ-2, более того, КМ-1 у ruRoBERTa-large значительно выше КМ-1 у fastText. В свою очередь, это может говорить о том, что модель семейства BERT лучше распознает словообразовательные признаки, чем fastText. С другой стороны, связь между словом и композицией его морфем в КМ-1, оцениваемая через косинусную метрику, не является сильной. Иными словами, мы не можем в этом случае ни подтвердить, ни опровергнуть гипотезу композициональности.

Табл. 2. Средние значения косинусной метрики для двух экспериментов.

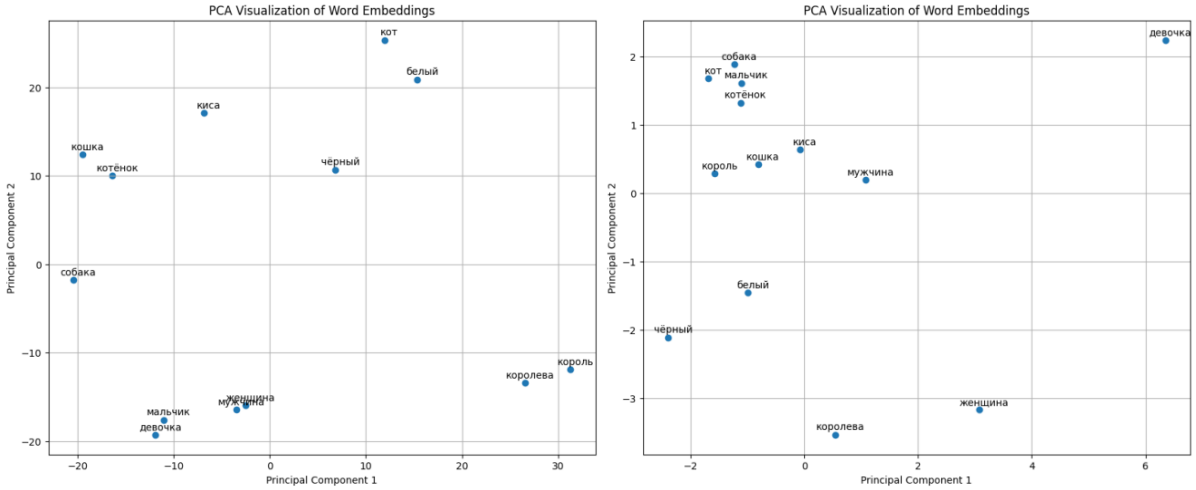
Модели		ВОДА	ЗЕМЛЯ	СВЕТ	ЛЕС	ДОМ	СЛОВО	СНЕГ
Объем гнезда		51	50	49	47	45	44	43
fastText	КМ-1, mean	0.423	0.526	0.456	0.485	0.46	0.479	0.512
geowac_lemma	КМ-2, mean	**0.797**	**0.71**	**0.757**	**0.723**	**0.69**	**0.668**	**0.743**
ruRoBERTa-large	КМ-1, mean	**0.646**	**0.708**	**0.696**	**0.707**	**0.703**	**0.682**	**0.741**
subword aggregation = mean	КМ-2, mean	0.387	0.371	0.384	0.379	0.332	0.331	0.364

Итак, результаты проведенного эксперимента подтверждают, что предсказывающие модели распределенных векторных вложений недостаточно полно воспроизводят синтаксические отношения между морфемами и поэтому не могут представлять композиционную семантику производных слов при агрегации

морфемных векторов. Таким образом, наша гипотеза не подтверждена. Полученные результаты стимулируют исследования, направленные на поиск нейросетевых архитектур, которые позволили бы обучить искомые модели. На возможность решения такой задачи указывает и то, что в языковых моделях воспроизводятся синтагматические отношения внутри предложений. Это означает, что при наличии соответствующей разметки на уровне морфемики и морфологии модели смогут интерпретировать подобные связи и внутри слова.

Оценка способности моделей воспроизводить семантические отношения внутри словообразовательных гнезд

Дополнительно был проведен второй эксперимент, направленный на сравнение моделей fastText и ruRoBERTa в задаче установления семантических связей между словами в словообразовательных гнездах с опорой на векторы подслов и морфем. Результаты представлены на рис. 3. Очевидно, что для модели fastText родовидовые отношения и дифференциация по признаку пола являются однонаправленными и более близкими, чем в модели ruRoBERTa (ср. векторы для лемм *кошка*, *котенок; кот* и *киса* более компактно расположены в fastText и более рассредоточены в пространстве ruRoBERTa). Значит, модель ruRoBERTa регистрирует семантическую близость векторов слов без учета их морфемного состава и словообразовательных отношений, тогда как векторы fastText передают информацию как о близости лексических значений слов, так и об их внутренней форме (в понимании А. А. Потебни).



(а) (б)

Рис. 3. Векторизация целевых слов: а) в ruRoBERTa; б) в fastText.

ЗАКЛЮЧЕНИЕ

В ходе исследования была предпринята попытка оценить способность русскоязычных языковых моделей воспроизводить синтагматические связи между морфемами внутри производных слов и на уровне словообразовательных гнезд. Основное внимание было сосредоточено на проверке гипотезы о композициональности производных слов при агрегации векторов морфем.

Эксперименты с моделями fastText и ruRoBERTa-large показали, что наилучшие результаты могут быть получены с использованием усреднения для агрегации векторов морфем, при этом сравнение вектора основы с разностью вектора слова и агрегированного вектора морфем демонстрирует более высокие значения, чем сравнение вектора слова с агрегированным вектором морфем.

Эксперимент по оценке способности моделей воспроизводить семантические отношения внутри словообразовательных гнезд показал, что модель fastText лучше передает информацию как о близости лексических значений слов, так и об их внутренней форме, в то время как модель ruRoBERTa-large регистрирует семантическую близость векторов слов без учета их морфемного состава и словообразовательных отношений.

Гипотеза о композициональности производных слов при агрегации морфемных векторов не получила однозначного подтверждения. Как показал эксперимент с моделями fastText, наилучшие результаты агрегации векторов морфем

достигаются с использованием усреднения, при этом сравнение вектора основы с разностью вектора слова и агрегированного вектора морфем дает более высокие значения близости, чем сравнение вектора слова с агрегированным вектором морфем. При оценке семантических связей слов внутри словообразовательных гнезд модель fastText подтвердила способность учитывать как близость значений слов, так и их словообразовательные связи, тогда как модель ruRoBERTa воспроизводит преимущественно лексико-семантические отношения.

Перспективы развития настоящего исследования связаны с разработкой специализированных нейросетевых архитектур, учитывающих синтагматические отношения между морфемными сегментами внутри слов, совершенствованием морфемных токенизаторов, интегрируемых в языковые модели, расширением наборов данных для решения вышеуказанных задач, а также с развитием комбинированных подходов, объединяющих преимущества моделей семейств fastText и BERT.

СПИСОК ЛИТЕРАТУРЫ

1. *Гердд А.С.* Морфемика. СПб.: Изд-во С.-Петерб. ун-та, 2004. 176 с.

2. *Bolshakova E.I., Sapin A.S.* Building a Combined Morphological Model for Russian Word Forms. In: Burnaev E. et al. Analysis of Images, Social Networks and Texts. AIST 2021. Lecture Notes in Computer Science, vol. 13217. Springer, Cham, 2022. P. 45–55. https://doi.org/10.1007/978-3-031-16500-9_5

3. *Bolshakova E.I., Sapin A.S.* Building Dataset and Morpheme Segmentation Model for Russian Word Forms. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference «Dialogue». Moscow, 2021. P. 154–161. https://doi.org/10.28995/2075-7182-2021-20-154-161

4. *Morozov D., Shcherbakova O., Glazkova A.* Russian Neural Morpheme Segmentation: From Lemmata to Wordforms. In: Bakaev M. et al. Internet and Modern Society. IMS 2025. Communications in Computer and Information Science, vol. 2671. Springer, Cham, 2025. https://doi.org/10.1007/978-3-032-04958-2_12, P. 157–167.

5. *Morozov D., Astapenka L., Glazkova A., Garipov T., Lyashevskaya O.* BERT-like Models for Slavic Morpheme Segmentation. In: Che W., Nabende J., Shutova E., Pilehvar M.T. (Eds.) Proceedings of the Annual Meeting of the Association

for Computational Linguistics. Association for Computational Linguistics, 2025. P. 6795–6815. (Proceedings of the Annual Meeting of the Association for Computational Linguistics). https://doi.org/10.18653/v1/2025.acl-long.337

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

7. *Selkirk E.* The syntax of words. Camb. (Mass), 1982. 136 p.

8. *Skalička V.* Hyposyntax. In: Slovo a slovesnost. Vol. 31. 1970. P. 1–6.

9. *Кубрякова Е.С.* Основы морфологического анализа. М., 1974. 320 с.

10. *Лопатин В.В.* Грамматическое описание славянских языков // Словообразование как объект грамматического описания. М., 1974.

11. *Lees R.* The Grammar of English nominalizations. The Hague, 1963.

12. *Marchand H.* The Categories and Types of Present-day English Word-Formation. Wiesbaden, 1960.

13. *Фивейская Е.А.* Словообразовательное моделирование семантики отглагольных имен в аспекте теории пропозиции // Сибирский филологический журнал. 2010 (3). С. 127–133.

14. *Филлмор Ч.* Дело о падеже // Новое в зарубежной лингвистике. Вып. 10. М., 1981.

15. *Шадрин В.И.* Семантика морфологических компонентов производных слов английского языка в свете категорий падежной грамматики // Морфемика. Принципы сегментации, отождествления и классификации морфологических единиц / Под ред. С.И. Богданова, А.С. Герда. СПб., 1997. С. 171–177.

16. *Morfessor.* URL: https://github.com/aalto-speech/morfessor, дата обращения 24.03.2026

17. *RussianMorphParsing.*
URL: https://github.com/alesapin/RussianMorphParsing, дата обращения 24.03.2026

18. *ruMorpheme*. URL: https://github.com/EvilFreelancer/ruMorpheme, дата обращения 24.03.2026

19. *Neuromodels.*
URL: https://ruscorpora.ru/license-content/neuromodels/, дата обращения

24.03.2026

20. *Asgari E., El Kheir Y., Sadraei Javaheri M. A.* MorphBPE: A Morpho-Aware Tokenizer Bridging Linguistic Complexity for Efficient LLM Training Across Morphologies, 2025. https://doi.org/10.48550/arXiv.2502.00894

21. *Teklehaymanot et al.* MoVoC: Morphology-Aware Subword Construction for Ge'ez Script Languages. In: Findings of the Association for Computational Linguistics: EMNLP 2025, p. 13131–13144, Suzhou, China. Association for Computational Linguistics, 2025. https://doi.org/10.48550/arXiv.2509.08812

22. *Nzeyimana A., Niyongabo Rubungo A.* KinyaBERT: a Morphology-aware Kinyarwanda Language Model. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), p. 5347–5363, Dublin, Ireland. Association for Computational Linguistics, 2022. https://doi.org/10.48550/arXiv.2203.08459

23. *Потиха З.А.* Школьный словарь строения слов русского языка: Пособие для учащихся. 2-е изд., испр. М.: Просвещение, 1999. 318 с.

24. *Тихонов А.Н.* Морфемно-орфографический словарь русского языка. М.: АСТ: Астрель, 2002. 704 с.

25. *Bojanowski P., Grave E., Joulin A., Mikolov T.* Enriching Word Vectors with Subword Information. In: Transactions of the Association for Computational Linguistics, 2017. P. 135-146. https://doi.org/10.48550/arXiv.2309.10931

26. *RusVectōrēs*. URL: https://rusvectores.org/ru/models/, дата обращения 24.03.2026

27. *Zmitrovich D., Abramov A., Kalmykov A., Kadulin V., Tikhonova M., Taktasheva E., Astafurov D., Baushenko M., Snegirev A., Shavrina T., Markov S., Mikhailov V., Fenogenova A.* 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). Torino, Italia, 2024. P. 507–524. https://doi.org/10.48550/arXiv.2309.10931

28. *ruRoBERTa-large.* URL: https://huggingface.co/ai-forever/ruRoBERTa-large, дата обращения 24.03.2026

REPRESENTATION OF INTRAWORD SYNTAGMATIC RELATIONS IN VECTOR LANGUAGE MODELS

D. K. Rodionova[1] [0009-0004-6296-8532], O. A. Mitrofanova[2] [0000-0002-3008-5514]

[1, 2]*Saint-Petersburg State University, Saint-Petersburg, Russia*

[1]*Chebyshev Research Center, Saint-Petersburg, Russia*

[1]rodionowadarja@yandex.ru, [2]o.mitrofanova@spbu.ru

Abstract

The paper discusses semantic structure representation of derivatives in language models, taking into account the intraword syntagmatic relations between derivational morphemes. Experiments were conducted using morphemic models developed by the Russian National Corpus (RNC), as well as fastText and ruRoBERTa models. The study is aimed at the verification of the hypothesis dealing with compositionality of derived words which are represented as aggregated morpheme vectors. In experiments we explore the representation of semantic relationships using fastText morpheme vectors and standard subword vectors in ruRoBERTa. The results indicate moderate sensitivity of fastText vectors to syntagmatic relations between morphemes as well as to derivational types. At the same time, it was found that aggregating morpheme vectors in fastText provides better representation of semantic relations between words compared to aggregating subword vectors in ruRoBERTa.

Standard BPE (Byte-Pair Encoding) and WordPiece tokenizers used in Transformer-based models are poorly interpretable with respect to linguistic data, as word segments do not always correspond to morphemes. The research problem lies in the need to assess the extent to which modern language models can capture linguistic features that characterize the relationships of derived words within word-formation families. The aim of the study is to evaluate the ability of predictive distributed vector embedding models to reproduce syntagmatic connections between morphemes within derived words and at the level of word-formation families in the Russian language.

The obtained results encourage the development of neural network architec-

tures that take into account syntagmatic relations between morphemes, the improvement of morpheme tokenizers, and their integration into language models.

***Keywords**: language models, morphemic analysis, word-formation methods, compositionality.*

REFERENCES

1. *Gerd A.S.* Morphology. St. Petersburg: Publishing House of St. Petersburg University, 2004. 176 p.

2. *Bolshakova E.I., Sapin A.S.* Building a Combined Morphological Model for Russian Word Forms. In: Burnaev, E., et al. Analysis of Images, Social Networks and Texts. AIST 2021. Lecture Notes in Computer Science, vol. 13217. Springer, Cham, 2022. P. 45–55. https://doi.org/10.1007/978-3-031-16500-9_5

3. *Bolshakova E.I., Sapin A.S.* Building Dataset and Morpheme Segmentation Model for Russian Word Forms. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”. Moscow, 2021. P. 154–161. https://doi.org/10.28995/2075-7182-2021-20-154-161

4. *Morozov D., Shcherbakova O., Glazkova A.* Russian Neural Morpheme Segmentation: From Lemmata to Wordforms. In: Bakaev M. et al. Internet and Modern Society. IMS 2025. Communications in Computer and Information Science, vol. 2671. Springer, Cham, 2025. P. 157–167.
https://doi.org/10.1007/978-3-032-04958-2_12

5. *Morozov D., Astapenka L., Glazkova A., Garipov T., Lyashevskaya O.* BERT-like Models for Slavic Morpheme Segmentation. In: Che W., Nabende J., Shutova E., Pilehvar M.T. (Eds.) Proceedings of the Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2025. P. 6795–6815 (Proceedings of the Annual Meeting of the Association for Computational Linguistics). https://doi.org/10.18653/v1/2025.acl-long.337

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

7. *Selkirk E.* The syntax of words. Camb. (Mass), 1982. 136 p.

8. *Skalička V.* Hyposyntax. In: Slovo a slovesnost. Vol. 31. 1970. P. 1–6.

9. *Kubryakova E.S.* Fundamentals of Morphological Analysis. Moscow, 1974. 320 p.

10. *Lopatin V.V.* Grammatical Description of Slavic Languages // Word Formation as an Object of Grammatical Description. Moscow, 1974.

11. *Lees R.* The Grammar of English nominalizations. The Hague, 1963.

12. *Marchand H.* The Categories and Types of Present-day English Word-Formation. Wiesbaden, 1960.

13. *Fiveyskaya E.A.* Word-Formation Modeling of the Semantics of Verbal Nouns in the Aspect of Proposition Theory // Siberian Philological Journal. 2010(3). P. 127–133.

14. *Fillmore C.* The Case for Case // New in Foreign Linguistics. Issue 10. Moscow, 1981.

15. *Shadrin V.I.* The Semantics of Morphological Components of Derived Words in the English Language in Light of the Categories of Case Grammar // Morphemics. Principles of Segmentation, Identification, and Classification of Morphological Units / Ed. by S.I. Bogdanov, A.S. Gerd. St. Petersburg, 1997. P. 171–177.

16. *Morfessor*. URL: https://github.com/aalto-speech/morfessor, last access 24.03.2026

17. *RussianMorphParsing.*
URL: https://github.com/alesapin/RussianMorphParsing, last access 24.03.2026

18. *ruMorpheme.* URL: https://github.com/EvilFreelancer/ruMorpheme, last access 24.03.2026

19. *Neuromodels.*
URL: https://ruscorpora.ru/license-content/neuromodels/, last access 24.03.2026

20. *Asgari E., El Kheir Y., Sadraei Javaheri M.A.* MorphBPE: A Morpho-Aware Tokenizer Bridging Linguistic Complexity for Efficient LLM Training Across Morphologies, 2025. https://doi.org/10.48550/arXiv.2502.00894

21. *Teklehaymanot et al.* MoVoC: Morphology-Aware Subword Construction for Ge'ez Script Languages. In: Findings of the Association for Computational Linguistics: EMNLP 2025, p. 13131–13144, Suzhou, China. Association for Computational Linguistics, 2025. https://doi.org/10.48550/arXiv.2509.08812

22. *Nzeyimana A., Niyongabo Rubungo A.* KinyaBERT: a Morphology-aware

Kinyarwanda Language Model. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), P. 5347–5363, Dublin, Ireland. Association for Computational Linguistics, 2022. https://doi.org/10.48550/arXiv.2203.08459

23. *Potikha Z.A.* School Dictionary of Word Structure of the Russian Language: A Guide for Students. 2nd ed., revised. Moscow: Prosveshchenie, 1999. 318 p.

24. *Tikhonov A.N.* Morphemic-Orthographic Dictionary of the Russian Language. Moscow: AST: Astrel, 2002. 704 p.

25. *.Bojanowski P., Grave E., Joulin A., Mikolov T.* Enriching Word Vectors with Subword Information. In: Transactions of the Association for Computational Linguistics, 2017. P. 135–146. https://doi.org/10.48550/arXiv.2309.10931

26. *RusVectōrēs.* URL: https://rusvectores.org/ru/models/, last access 24.03.2026

27. *Zmitrovich D., Abramov A., Kalmykov A., Kadulin V., Tikhonova M., Taktasheva E., Astafurov D., Baushenko M., Snegirev A., Shavrina T., Markov S., Mikhailov V., Fenogenova A.* 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). Torino, Italia, 2024. P. 507–524. https://doi.org/10.48550/arXiv.2309.10931

28. *ruRoBERTa-large.*
URL: https://huggingface.co/ai-forever/ruRoBERTa-large, last access 24.03.2026

СВЕДЕНИЯ ОБ АВТОРАХ

РОДИОНОВА Дарья Кирилловна – магистрант кафедры математической лингвистики филологического факультета Санкт-Петербургского государственного университета, старший инженер-программист Chebyshev Research Center. В 2014 году закончила бакалавриат на кафедре информационных систем в области искусств и гуманитарных наук факультета искусств Санкт-Петербургского государственного университета. В 2018 году окончила обучение Computer Science Center при поддержке компании JetBrains, слушала курсы ШАДа. Основные научные интересы связаны с языковым моделированием, математической лингвистикой, информационным поиском, извлечением знаний, анализом кода методами NLP и машинным обучением.

Daria Kirillovna RODIONOVA – master student at the Department of Mathematical Linguistics, Faculty of Philology, Saint-Petersburg State University, and senior software engineer at Chebyshev Research Center. In 2014 she graduated with a Bachelor degree from the Department of Information Systems in the Arts and Humanities at the Faculty of Arts of Saint-Petersburg State University. In 2018 she completed training at Computer Science Center supported by JetBrains and attending courses at the Yandex School of Data Analysis. Her main scientific interests are related to language modeling, mathematical linguistics, information retrieval, knowledge extraction, code analysis using NLP methods, and machine learning.

email: rodionowadarja@yandex.ru

ORCID: 0009-0004-6296-8532

МИТРОФАНОВА Ольга Александровна – кандидат филологических наук, доцент кафедры математической лингвистики филологического факультета Санкт-Петербургского государственного университета. В 1995 году закончила отделение математической лингвистики филологического факультета Санкт-Петербургского государственного университета, в 1999 году защитила диссертацию на соискание ученой степени кандидата филологических наук по специальности 10.02.21 – Прикладная и математическая лингвистика. Является автором более 150 публикаций в области компьютерной и корпусной лингвистики. Основные научные интересы связаны с моделями языка, машинным обучением, автоматическим пониманием и генерацией текстов, лингвистикой конструкций, дистрибутивной семантикой, тематическим моделированием.

Olga Aleksandrovna MITROFANOVA – PhD in Philology, associate professor at the Department of Mathematical Linguistics, Faculty of Philology, Saint-Petersburg State University. She graduated from Mathematical Linguistics Department, Faculty of Philology, Saint-Petersburg State University in 1995, and in 1999 she defended her thesis in Applied and Mathematical Linguistics (10.02.21). She is the author of over 150 publications in the field of Computational and Corpus Linguistics. Her main research interests are language models, machine learning, natural text understanding and generation, construction linguistics, distributional semantics, and topic modeling.

email: o.mitrofanova@spbu.ru

ORCID: 0000-0002-3008-5514

Материал поступил в редакцию 23 марта 2026 года