DOI: 10.15514/ISPRAS-2021-33(4)-9

Построение нейросетевых моделей морфологического и морфемного анализа текста

А.С. Сапин, ORCID: 0000-0002-9532-132X<alesapin@gmail.com>
Московский государственный университет имени М.В. Ломоносова,
119991, Россия, Москва, Ленинские горы, д. 1

Аннотация. Морфологический анализ текстов на естественном языке является одним из важнейших этапов автоматической обработки текстов (АОТ). Традиционные и хорошо исследованные задачи морфологического анализа включают приведение словоформы к нормальной форме (лемме), определение ее морфологических характеристик, а также разрешение (снятие) морфологической омонимии (неоднозначности характеристик). К морфологическому анализу относится также задача морфемного разбора слов (т.е. сегментация слов на составляющие морфы и их классификация), которая востребована в некоторых приложениях АОТ. В последние годы разработан ряд программных моделей на основе машинного обучения, повышающих точность традиционного морфологического анализа и морфемного разбора, однако производительность таких моделей недостаточна для многих практических задач, а для задачи морфемного разбора высокоточные модели построены только для лемм. В данной работе описаны две новые высокоточные нейросетевые модели, реализующие морфемный разбор словоформ русского языка при достаточно высокой производительности. Первая модель основана на сверточной нейронной сети и показывает достойное качество морфемного разбора словоформ. Вторая модель, кроме морфемного разбора словоформы, позволяет предварительно уточнить её морфологические характеристики, решая задачу снятия омонимии. Производительность этой объединенной морфологической модели оказалась наилучшей среди рассмотренных моделей морфемного разбора, при сравнимой точности разбора.

Ключевые слова: морфологический анализ словоформ; автоматический морфемный разбор; нейросетевые модели морфемного разбора.

Для цитирования: Сапин А.С. Построение нейросетевых моделей морфологического и морфемного анализа текста. Труды ИСП РАН, том 33, вып. 4, 2021 г., стр. 117-130. DOI: 10.15514/ISPRAS–2021–33(4)–9

Building neural network models for morphological and morpheme analysis of texts

A.S. Sapin, ORCID: 0000-0002-9532-132X <alesapin@gmail.com>
Lomonosov Moscow State University,
GSP-1, Leninskie Gory, Moscow, 119991, Russia

Abstract. Morphological analysis of text is one of the most important stages of natural language processing (NLP). Traditional and well-studied problems of morphological analysis include normalization (lemmatization) of a given word form, recognition of its morphological characteristics and their morphological disambiguation. The morphological analysis also involves the problem of morpheme segmentation of words (i.e., segmentation of words into constituent morphs and their classification), which is actual in some NLP applications. In recent years, several machine learning models have been developed, which increase the accuracy of traditional

morphological analysis and morpheme segmentation, but performance of such models is insufficient for many applied problems. For morpheme segmentation, high-precision models have been built only for lemmas (normalized word forms). This paper describes two new high-accuracy neural network models that implement morphemic segmentation of Russian word forms with sufficiently high performance. The first model is based on convolutional neural networks and shows the state-of-the-art quality of morphemic segmentation for Russian word forms. The second model, besides morpheme segmentation of a word form, preliminarily refines its morphological characteristics, thereby performing their disambiguation. The performance of this joined morphological model is the best among the considered morpheme segmentation models, with comparable accuracy of segmentation.

Keywords: morpheme segmentation of wordforms; neural models for morphological analysis; morphological analysis of wordforms

For citation: Sapin A.S. Building neural network models for morphological and morpheme analysis of texts. Trudy ISP RAN/Proc. ISP RAS, vol. 33, issue 4, 2021, pp. 117-130 (in Russian). DOI: 10.15514/ISPRAS-2021-33(4)-9

1. Введение

Морфологический анализ является одним из базовых этапов автоматической обработки текстов (АОТ), результаты которого используются во многих прикладных задачах. К основным задачам морфологического анализа относится определение морфологических характеристик (часть речи, падеж, число, род и т.д.) словоформы [1]. Например, для словоформы “шоколада” распознаются характеристики: существительное, родительного падежа, единственного числа, мужского рода.

Важной задачей морфологического анализа является снятие (разрешение) морфологической неоднозначности (омонимии), т.е. выявление корректного для обрабатываемого текста варианта *морфологических характеристик* словоформы из всех возможных. Например, словоформа *“стали”* может быть как существительным множественного числа (“*виды стали”*), так и глаголом прошедшего времени (“*стали разгружать”*). Разрешение омонимии в этом примере сводится к выбору одного варианта из двух возможных *<сущ., мн. ч., …>* и *<гл., пр. вр., …>*. Качество морфологического анализа обычно оценивается с учетом снятия омонимии, для этого используется метрика аккуратности (точности) определения морфологических характеристик [2], которая рассчитывается как количество правильных ответов к количеству всех анализируемых словоформ текста.

Ещё одной задачей, относящейся к морфологическому анализу, является морфемный разбор [3], который заключается в анализе морфемного состава слова путем его разбиения (сегментации) на морфы (морфемы), например: *impossible → im-poss-ible*, *прекрасный → пре-крас-н-ый*. Морфемы являются наименьшими значащими единицами текста, и результаты морфемного разбора необходимы в ряде прикладных задач АОТ, таких как исправление словообразовательных и паронимических ошибок, распознавание смысла незнакомых и редких слов по более частотным родственным словам.

Задачи разрешения морфологической омонимии и морфемного разбора являются актуальными для высокофлективных языков со сложным словоизменением и словообразованием (большое количество суффиксов, префиксов, окончаний), к каковым относится русский язык. В последние годы продолжаются исследования по применению машинного обучения для задач морфологического анализа русского языка [4, 5, 6, 7], которые позволили улучшить качество разрешения морфологической омонимии до 95% точности для морфологических характеристик. Однако производительность таких машинообученных моделей анализа является чрезвычайно низкой (всего лишь сотни слов в секунду на одном ядре CPU). Для задачи морфемного разбора на базе машинного обучения были построены высокоточные модели разбора лемм (нормальных форм) русского языка [8, 9, 10], однако их точность для различных словоформ русского языка недостаточна и их производительность не оценивалась.

Настоящая работа посвящена проблеме эффективности программных моделей морфологического анализа, в том числе морфемного разбора, для словоформ русского языка. Под эффективностью мы понимаем как высокую точность решения задач морфологического анализа, так и производительность, позволяющую быстрее обрабатывать современные корпуса текстов из сотен миллионов слов.

В работе реализованы и исследованы две новые нейросетевые модели морфемного разбора словоформ русского языка. Первая модель выполняет морфемный разбор словоформ, превосходя как по производительности, так и по точности разбора известные модели морфемного разбора для лемм [8, 9, 10]. Поскольку для применения этой модели необходима такая морфологическая характеристика словоформы, как часть речи, дополнительно на основе этой модели построена вторая, объединенная модель морфологического анализа, выполняющая одновременно снятие морфологической омонимии словоформы и её морфемный разбор.

В следующем разделе кратко излагаются результаты в области традиционного морфологического анализа, применимые для построения процессоров русского языка. В третьем разделе рассматриваются подходы к автоматическому морфемному разбору и разработанные в последние годы высокоточные модели морфемного разбора лемм русского языка. В четвертом разделе описывается разработанная нами модель морфемного разбора словоформ русского языка, а также её экспериментальное исследование. В пятом разделе содержится описание объединенной модели морфологического анализа: ее архитектура, параметры обучения, оценки качества и производительности. В заключении кратко приводятся основные результаты настоящей работы.

2. Методы традиционного морфологического анализа

Для русского языка большинство применяемых в настоящее время морфологических процессоров (в том числе открытые анализаторы [11, 12, 13]) базируются на словарной информации, т.е. либо на словаре основ, либо на словаре словоформ (последние для русского языка используются значительно чаще). Определение морфологических характеристик анализируемой словоформы сводится к её поиску в соответствующем словаре и выдаче всех возможных вариантов морфологических характеристик (тегов) обрабатываемой словоформы.

Словарные морфологии показывают высокую производительность (до 120 тысяч слов в секунду на CPU [11]), однако не позволяют решать задачу снятия морфологической омонимии. Для её решения требуется последующее применение отдельной процедуры к полученным из словаря результатам. Эта процедура обычно строится на основе машинного обучения с учителем по размеченному текстовому корпусу и позволяет выбрать единственно верный вариант морфологических характеристик из нескольких возможных. В разных морфологических процессорах используются разные методы машинного обучения: в Диалинг-АОТ [11] – скрытые марковские цепи, в TreeTagger [14] – деревья решений, а в парсере UDPipe 1.0 [15] – полносвязная нейронная сеть. Подобные методы достигают точности определения морфологических характеристик с учетом снятия морфологической омонимии до 94.5% для известных слов и до 79% для слов отсутствующих в словарях [2]. Точность (аккуратность) определения морфологических характеристик рассчитывается как отношение количества словоформ, у которых характеристики определены верно, к количеству всех анализируемых словоформ:

$$A_{tags} = \frac{\sum_{i=0}^{len(dataset)} correct(word_i)}{len(dataset)},$$

где *len(dataset)* – количество словоформ в анализируемом тексте, $word_i$ – *i*-ое слово в тексте, а $correct(word_i) = 1$, когда все морфологические характеристики слова определены верно, и равно 0 в противном случае.

В последние годы были предложены модели морфологического анализа, в которых определение морфологических характеристик и снятие омонимии происходит одновременно, т.е. для каждой словоформы сразу же находится единственный вариант леммы и морфологических характеристик [4, 5, 6]. Особенностью такого подхода является использование векторных представлений слов из нейронных языковых моделей разного вида: FastText [16], ELmO [17], BERT [18]. В работе [5] использовались контекстуализированные векторные представления BERT и мультиклассовая логистическая регрессия и было достигнуто наилучшее качество решения задач морфологического анализа для русского языка: 95% точности определения морфологических характеристик. Такие показатели качества достаточны для прикладных задач АОТ, однако производительность подобных высокоточных моделей оказывается более чем на два порядка ниже словарных методов, поэтому их применение в практических приложениях ограничено. Открытые морфологические процессоры русского языка [11, 12, 13] по-прежнему базируются на словарях и более простых методах снятия морфологической омонимии.

3. Методы морфемного разбора лемм

Известны два варианта морфемного разбора слов:

- *морфемная сегментация*, когда требуется сегментировать слово на составляющие его морфы (морфемы), например, для слова *сетка* – *сет-к-а*;
- морфемная *сегментация с классификацией*, когда требуется не только сегментировать слово на морфы, но и определить их тип: приставка (*PREF*), корень (*ROOT*), суффикс (*SUFF*), окончание (*END*) и т.д., например, *сетка* – *сет:ROOT/к:SUFF/а:END*.

Морфемная сегментация с классификацией является наиболее полным вариантом задачи морфемного разбора и именно она рассматривается в настоящей работе.

Качество автоматической морфемной сегментации оценивается с помощью метрик точности (*Presicion*), полноты (*Recall*) и F-меры по границам морфем [19], рассчитываемых следующим образом:

$$Precision = \frac{TP}{TP + FP};\ Recall = \frac{TP}{TP + FN};\ F = \frac{2TP}{2TP + (FP + FN)},$$

где *TP* – количество верно обнаруженных границ между морфами, *FP* – количество ложно обнаруженных границ, *FN* – количество не обнаруженных границ. Для задачи сегментации с классификацией добавляются ещё точность (аккуратность) определения типа всех получившихся морфем в сегментированном слове (аккуратность по словам целиком):

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

где, *len(dataset)* – количество словоформ в анализируемом тексте, $word_i$ – *i*-ое слово в тексте, а $correct(word) = 1$ только когда типы и границы всех морфов слова определены верно, и равно 0 иначе.

Первые методы автоматической морфемной сегментации [3] были чисто статистическими, основанными на неразмеченных данных и показывали 50-65% значения F-меры обнаружения границ морфем. Наиболее известное решение задачи морфемной сегментации было реализовано в системе Morfessor [20] на основе метода машинного обучения без учителя по большой неразмеченной коллекции текстов. Основная идея метода Morfessor состоит в поиске минимального набора морфем, с помощью которого можно сегментировать все слова обрабатываемой коллекции текстов. Для таких языков как английский, финский и турецкий система показывает около 70-80% F-меры для обнаруженных границ морфем.

Для применения машинного обучения с учителем нужны представительные наборы размеченных данных (*датасеты*) с сегментированными морфемами, но они трудоемки в создании и отсутствуют для большинства языков. Относительно недавно появились

несколько датасетов с морфемной разметкой (сегментация и классификация) для русского языка, наиболее представительный из них, RuMorphs-Lemmas[1], был получен на основе словообразовательного словаря Тихонова [21] и содержит около 96 тысяч размеченных лемм русского языка. Благодаря этому, на основе методов машинного обучения с учителем были разработаны несколько высокоточных методов (моделей) морфемной сегментации с классификацией для лемм русского языка [8, 9, 10]. В этих моделях использовались различные методы машинного обучения:

- сверточная нейронная сеть (CNN) [8];
- деревья решений с градиентным бустингом (GBDT) [9];
- двунаправленная нейронная LSTM-сеть (Bi-LSTM) [10].

Во всех моделях задача морфемного разбора рассматривалась как задача классификации букв, и помимо различий в методах машинного обучения модели различаются набором классов букв. CNN-модель применяет схему классификации BMES (используемую обычно в задаче выявления именованных сущностей), классифицируя каждую букву на 22 различных класса, а модели GBDT и Bi-LSTM используют сокращенный набор из 10 классов, но достаточный для решения рассматриваемой задачи. Во всех моделях буквы слова представляются в унитарной кодировке (*one-hot encoding*), а также учитывается информация о их гласности. Дополнительно, GBDT-модель использует значения морфологических характеристик сегментируемого слова: часть речи, род, число, падеж, время. Модель на основе двунаправленной LSTM-сети также применяет морфологическую информацию, но только часть речи. Важной особенностью CNN-модели является дополнительная корректирующая процедура на основе простых правил морфотактики (корень идет после приставки, суффикс после корня и т.п.), применяемая к результату нейронной сети, а также использование ансамбля из трех одинаковых CNN-моделей, что значительно повышает точность разбора, но увеличивает размер модели и снижает производительность.

Экспериментальная оценка [9] трёх указанных моделей на одних и тех же размеченных датасетах для русского языка (в том числе RuMorphs-Lemmas) показала их сравнимое качество: до 98-99% F-меры по границам морфем (в зависимости от обучающего датасета и параметров модели), а также 86-89% точности (аккуратности) морфемного разбора слов целиком – см. табл. 1 с оценками, полученными на датасете RuMorphs-Lemmas. Модель на основе Bi-LSTM слегка превосходит CNN-модель, возможно за счет дополнительного использования части речи разбираемого слова и сокращенного набора классов букв. В тоже время эта модель не требуют корректирующей процедуры. В работе [9] также показано, что наибольшее влияние на распознавание класса буквы оказывают не только соседние буквы, но и часть речи.

Табл. 1. Качество морфемного разбора для лемм русского языка (%)
Table 1. Quality of morphemic segmentation for Russian lemmas (%)

Модель	F-мера по границам морфем	Точность разбора слов
CNN + корректирующая процедура + ансамбль	98.10	88.62
GBDT + морфохарактеристики	98.01	86.54
Bi-LSTM + часть речи	**98.45**	**89.03**

Описанные модели морфемной сегментации с классификацией показывают высокую точность разбора лемм, однако их производительность не измерялась. Поскольку код моделей является открытым, мы произвели замеры их производительности на фрагменте

[1] https://github.com/cmc-msu-ai/NLPDatasets/blob/main/morphemic/dicts/tikhonov.txt

коллекции текстов lib.rus.ec[2], объемом 10 млн слов, в одноядерном режиме процессора Intel Core I7-8750H, без использования графического ускорителя. Измерялось количество слов в секунду, обрабатываемых моделью (с учетом времени на определение морфологических характеристик в моделях GBDT и Bi-LSTM и корректирующей процедуры в CNN-модели), результаты показаны в табл. 2.

Производительность моделей оказалась невысока, так что для обработки большой коллекции текстов (сотни миллионов слов), даже с учетом параллелизма потребуется несколько дней. Наибольшая производительность достигается моделью, построенной на базе сверточных нейронных сетей, она же имеет и наименьший размер.

Табл. 2. Производительность моделей морфемного разбора лемм русского языка
Table 2. Performance morphemic segmentation of Russian lemmas

Модель для лемм	Слов в секунду	Размер модели (МБ)
CNN + корректирующая процедура + ансамбль	**354**	**9.5**
GBDT + определение морфохарактеристик	269	2651
Bi-LSTM + определение части речи	64	203

Отметим, что все описанные выше модели морфемного разбора были обучены для морфемного разбора лемм (нормальных форм) русского языка, и выполненные замеры качества морфемного разбора для словоформ показали их непригодность для практики (менее 38% точности разбора по словам целиком). Причиной этого является существенное различие в морфемной структуре различных словоформ морфологически богатого русского языка, например:

разбор леммы: *расшить – рас:PREF/ши:ROOT/ть:END*
разбор словоформы: *разошьют – разо:PREF/шь:ROOT/ют:END*

Поскольку тексты состоят не из лемм, а из словоформ, необходима эффективная модель морфемного разбора, ориентированная на обработку словоформ.

4. Сверточная модель морфемного разбора словоформ

Известно, что сверточные нейронные сети являются одним из наиболее производительных видов нейронных сетей, как для обучения, так и для применения уже обученной модели. При сравнении моделей морфемного разбора лемм (табл. 2) модель на базе сверточных нейронных сетей также показала наилучшую производительность. Поэтому мы выбрали одномерные сверточные нейронные сети в качестве основы архитектуры модели морфемного разбора словоформ. Поскольку сверточные сети работают с последовательностями фиксированной длины, наша модель обрабатывает слова из 20 букв (подавляющее число слов русского языка содержит меньшее число букв). Более короткие слова дополняются пустыми символами, а более длинные делятся на части. Архитектура сети представлена на рис. 1.

На вход разработанной CNN-модели подается числовой вектор из закодированных букв словоформы в унитарной кодировке, признаков их гласности, а также закодированной части речи словоформы. Вход модели соединен со “сверточным блоком”, который состоит из одномерного сверточного слоя, слоя субдискретизации (*max pooling*) и слоя исключения (*dropout*). Слой субдискретизации позволяет значительно ускорить обучение и последующее применение модели, а слой исключения помогает бороться с переобучением. В качестве функции активации сверточного слоя взята ReLU, которая является одновременно вычислительно простой и хорошо зарекомендовавшей себя на практике. Всего в модели

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

используются три последовательно соединённых “сверточных блока”, выход последнего подается на вход полносвязным слоям сети с функцией активации мягкий максимум (*softmax*), выступающим в роли классификаторов. Эксперименты показали, что увеличение числа сверточных блоков незначительно улучшает качество разбора, но снижает производительность модели.

Поскольку для обучения модели морфемного разбора словоформ необходим размеченный датасет с морфемным разбором словоформ (а не лемм), нами была разработана автоматическая процедура генерации размеченного датасета словоформ, исходя из известного датасета RuMorphs-Lemmas. Процедура последовательно принимает на вход морфемный разбор очередной леммы русского языка из этого датасета и на основе системы словоизменительных классов для русского языка и информации из морфологического словаря Opencorpora [22] генерирует разборы всех словоформ входной леммы. Построенный датасет RuMorphs-Words содержит более 1.7 млн различных словоформ с морфемной разметкой, для каждой словоформы указана её часть речи.

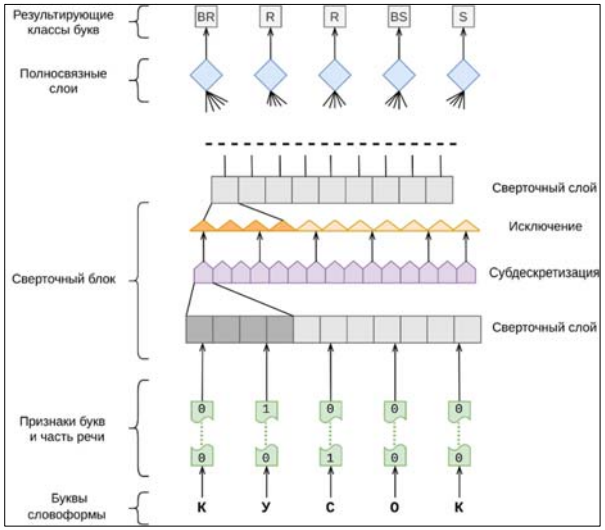

Рис. 1. Архитектура модели морфемного разбора словоформ
Fig 1. The architecture of the morphemic segmentation model of wordforms

При обучении модели буквы словоформы классифицируются на 10 классов, что достаточно для выделения соседних морфем, относящихся к одному и тому же типу (*ROOT*, *PREFIX*, *SUFFIX*). Ниже приведен пример, показывающий отличия более традиционной BMES-разметки (22 класса) от используемой нами BM-разметки (10 классов) на примере разбора словоформы *“мечтателя”, мечт:ROOT/а:SUFF/тел:SUFF/я:END*:

м	*е*	*ч*	*т*	*а*	*т*	*е*	*л*	*я*
B-ROOT	M-ROOT	M-ROOT	E-ROOT	S-SUFF	B-SUFF	M-SUFF	E-SUFF	S-END
B-ROOT	M-ROOT	M-ROOT	M-ROOT	B-SUFF	B-SUFF	M-SUFF	M-SUFF	B-END

Как видно, BM-разметка (нижняя строка) позволяет выделить границу последовательных суффиксов *“а”* и *“тель”*.

При обучении модели датасет разбивался в соотношении 70% для обучающего множества, 10% для валидационного и 20% для тестового (время обучения составило около 25 минут на Nvidia Tesla T4). Точность обученной модели морфемного разбора словоформ составила 91.06% по словам целиком для словоформ, а при проверке только на леммах – 90.03%, что является наилучшим достижимым качеством морфемного разбора для слов русского языка –

см. табл. 3, строка 1 (точность F-меры по границам морфем также высока, как и в моделях для разбора лемм, поэтому не показана).

Табл. 3. Точность моделей морфемного разбора словоформ русского языка (%)
Table 3. Accuracy of models for morphemic segmentation of Russian word forms (%)

Модель для словоформ	**RuMorphs-Words**	**RuMorphs-Lemmas**	**Morphs-SynTagRus**
CNN	90.03	91.06	-
Объединенная	-	85.90	88.54

Оценка производительности модели для словоформ выполнялась с помощью библиотеки tensorflow-lite [23], так как она включает большинство реализованных в tensorflow оптимизаций, а также обладает простым интерфейсом для применения моделей и поддерживается для нескольких языков программирования. Разработанная для словоформ модель показала наилучшую производительность среди рассмотренных моделей морфемного разбора: 4559 слов в секунду – см. табл. 4, строка 1. Однако, с учетом времени определения части речи (морфопроцессором[3] для русского языка) производительность снизилась до 2380 слов (строка 2 табл. 4). Тем самым, определение части речи негативно сказывается на производительности модели. Разработанная нами объединённая модель морфологического анализа позволяет добиться большей производительности за счёт одновременного определения части речи и морфемного разбора словоформы.

Табл. 4. Производительность моделей морфемного разбора словоформ
Table 4. Performance of models for morphemic segmentation of Russian word forms

Модель для словоформ	**Слов в секунду**	**Размер (МБ)**
CNN с известной частью речи	4559	1.1
CNN с определением части речи	2380	1.1
Объединенная морфологическая модель	1893	1.5
Комплекс объединенных морфологических моделей	3543	13.5

5. Объединенная модель морфологического анализа

Объединенная модель, так же, как и модель морфемного разбора словоформ, основана на сверточных нейронных сетях из-за их высокой производительности. В отличие от описанной выше CNN-модели для словоформ, объединённая модель обрабатывает текст по предложениям, последовательностям слов фиксированного размера.

Для каждой словоформы предложения берутся её возможные морфологические характеристики (варианты морфологического анализа), определяемые морфологическим процессором. В случае морфологической омонимии модель снимает её (уточняет часть речи, падеж, число, род, время) и использует уточненную часть речи для выполнения морфемного разбора.

Архитектура объединенной модели представлена на рис. 2, слева показана часть модели, отвечающая за разрешение морфологической омонимии, а справа – часть модели, отвечающая за морфемный разбор.

Поскольку использование векторных представлений слов, полученных из нейронных языковых моделей, значительно повышает качество морфологического анализа [4, 5, 6], на вход модели подаются вектора обрабатываемых словоформ из языковой модели FastText [16] (эта одна из вычислительно-простых языковых моделей для высокофлективного русского

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

языка). Эти вектора словоформ конкатенируются с векторами закодированных вариантов их морфологического анализа, полученными морфопроцессором.

Эти данные обрабатываются тремя сверточными блоками (их архитектура аналогична сверточным блокам вышеописанной модели морфемного разбора словоформ), полученный результат поступает в полносвязные слои (для каждого слова свой набор слоёв), выступающие в роли классификаторов и определяющие значения морфологических характеристик словоформ: часть речи, падеж, род, число, время.

Один выход полносвязного слоя для каждого слова, ответственный за часть речи, подается в ту часть модели, которая реализует морфемный разбор, вместе с закодированными буквами обрабатываемых словоформ и признаками их гласности (аналогично сверточной модели морфемного разбора словоформ). Морфемный разбор словоформ из обрабатываемой последовательности слов выполняется независимо.

Поскольку для обучения разрабатываемой модели необходим размеченный датасет, в котором будет одновременно и морфологическая, и морфемная разметка словоформ русского языка, а такие датасеты на данный момент не разработаны, то был взят и дополнительно размечен известный корпус с морфологической разметкой SynTagRus [24] (около 1.1 млн слов) – в нем была добавлена морфемная разметка каждой словоформы. Корпус SynTagRus был выбран, как представительный и в тоже время использованный в морфологическом соревновании [7], что позволяет сравнить разработанную нами модель с наилучшим достижимым качеством морфологического анализа. Морфемная разметка добавлялась в автоматизированном режиме с помощью нашей уже реализованной сверточной модели морфемного разбора словоформ и дополнительной ручной проверки результата.

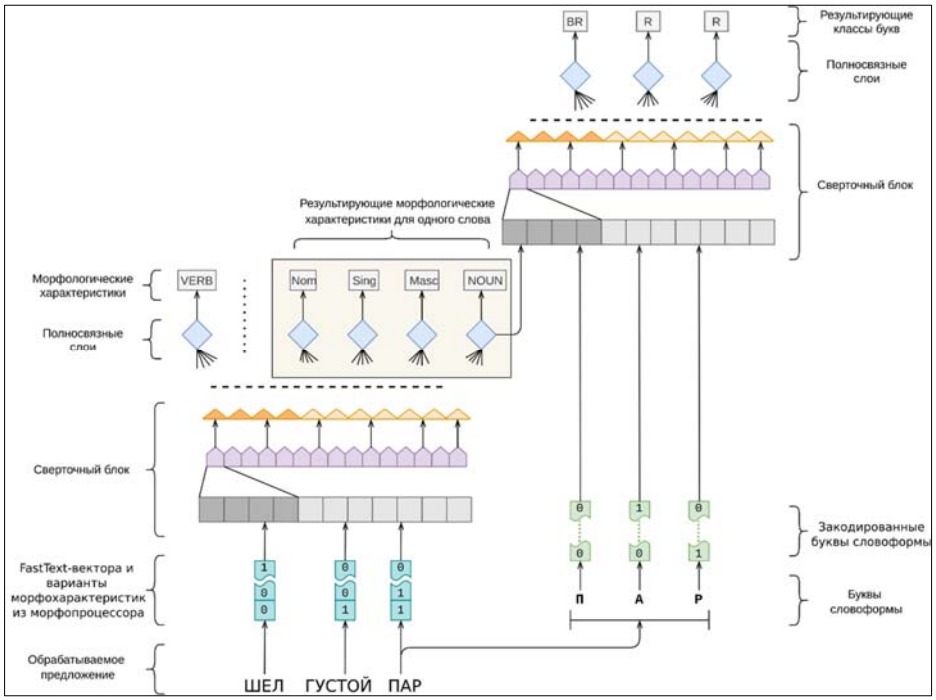

Рис. 2. Архитектура объединенной модели
Fig. 2. The architecture of the joined model

При обучении рассматриваемой объединенной модели использовалось следующее разбиение корпуса SynTagRus с морфемной разметкой, далее – Morphs-SynTagRus: 70% предложений для обучающего множества, 10% для валидационного и 20% для тестового. В ходе экспериментов с моделью было выяснено, что наилучшее качество морфологического анализа и морфемного разбора словоформ достигается при следующих гиперпараметрах: количество узлов в сверточных слоях равно соответственно 512, 256, 192, алгоритм градиентного спуска – Adam, со скоростью обучения равной 0.001. Величина исключения равна 0.3, а размер субдискретизации равен трём. Обучение такой модели занимает около 20 минут на видеокарте Nvidia Tesla T4.

Оценка модели, обученной для входных последовательностей из 9 слов, показала, что точность разрешения омонимии равна 94.2%, что несколько ниже наилучшего достижимого качества (96.5% [5]), а точность морфемного разбора по словам целиком достигает 96.5%, что значительно превосходит все предыдущие модели морфемного разбора. Заметим, что при оценке качества морфемного разбора не учитывались все слова из тестового множества короче трех букв, т.к. морфемный разбор таких слов тривиален, и оценка модели оказалась бы завышена. Однако при дополнительной валидации на датасете RuMorphs-Words модель показала значительно худший результат – 47.3% точности морфемного разбора целиком. Обнаруженная чрезмерная настройка модели на корпус Morphs-SynTagRus с морфемной разметкой объясняется в первую очередь тем, что слова в этом корпусе обладают очень низким “морфемным разнообразием”: в нем маленькое количество различных слов, большое количество коротких слов, в том числе повторяющихся или очень похожих по структуре.

Для преодоления обнаруженного недостатка был применен техника “переноса знаний” (*transfer learning*), часто используемая при создании нейронных моделей для обработки текстов. Обучение объединенной модели было разделено на 3 этапа. На первом этапе часть модели, отвечающая за морфемный разбор, обучалась отдельно на датасете RuMorphs-Words (с уже известными частями речи). На втором этапе веса в этой части нейронной модели замораживалась (т.е. исключались из обучения) и производилось обучение объединенной модели на доразмеченном корпусе Morphs-SynTagRus. На третьем этапе веса морфемной подмодели размораживались, скорость обучения устанавливалась на 2 порядка меньше, чем на втором этапе (для того, чтобы не потерять знания, полученные на этапе 1), и обучение всей объединенной модели производилось еще раз с максимальным количеством итераций равным 20 (по той же самой причине).

Таким образом, модель сохраняла знания о морфемных разборах, полученные на первом этапе обучения, и в тоже время обучалась разбирать словоформы из Morphs-SynTagRus. Это позволило добиться точности морфемного разбора 88.5% на словах из Morphs-SynTagRus и 85.9% для словоформ из RuMorphs-Words – см. табл. 3, строка 2. Последний показатель ниже наилучшего достижимого, однако заметим, что при уменьшении числа итераций на третьем этапе обучения точность морфемного разбора словоформ из RuMorphs-Words была более высокой, но при этом для Morphs-SynTagRus была ниже. Тем самым, изменяя количество итераций на третьем этапе обучения, модель можно настраивать на специфику одного или другого датасета.

Итоговое сравнение качества наилучшей модели для лемм (Bi-LSTM), CNN-модели для словоформ и объединённой модели по метрики точности сегментации с классификацией по словам целиком – см. табл. 5.

Табл. 5. Точность моделей морфемного разбора для русского языка (%)
Table 5. Accuracy of models for morphemic segmentation of Russian (%)

Модель	RuMorphs-Lemmas	RuMorphs-Words	Morphs-SynTagRus
Bi-LSTM (леммы)	89.03	38.57	34.49
CNN (словоформы)	90.03	91.06	-
Объединенная	85.11	85.90	88.54

Как видно из таблицы, модель Bi-LSTM для разбора лемм показывает плохое качество разбора словоформ. CNN-модель разбора словоформ показывает наилучшее достижимое качество на датасетах RuMorphs-Words и RuMorphs-Lemmas (на датасете Morphs-SynTagRus модель не оценивалась, так как с её помощью производилась разметка этого датасета). Объединенная морфологическая модель проигрывает по точности CNN-модели, хотя и не критично, однако её применение позволяет получить лучшую производительность.

Для тестирования производительности объединенной модели использовалась библиотека tensorflow-lite. Производительность модели оказалась равна 1893 слова в секунду – см. табл. 4, строка 3, что сравнимо с моделью морфемного разбора словоформ с учетом времени, затрачиваемого на определение части. Размер обученной объединенной модели составляет менее 1.5 мегабайт.

Описанная объединенная модель обучалась на входных последовательностях из девяти слов, до двадцати букв каждое. Поскольку в текстах часто встречаются короткие предложения, а также короткие слова, то при их обработке выполняются излишние вычисления (для дополненных до фиксированного размера концов таких предложений и слов). Для улучшения производительности предлагается использовать комплекс из 9 аналогичных объединённых моделей, для меньших размеров входных данных: 9 слов, 7 слов, 5 слов и, соответственно каждая из них для слов из 20 букв, 12 букв и 6 букв. Суммарный объем комплекса моделей составил около 13.5 мегабайт. При обработке входного текста делается выбор подходящей модели комплекса, т.е. размер слов в которой больше, чем во входном предложении, и количество букв в словах больше, чем у самого длинного слова. В этом случае производительность такого комплекса составила около 3543 слов в секунду, что является наилучшим результатом для морфемного разбора словоформ (табл. 4, строка 4).

6. Заключение

Разработаны и экспериментально исследованы две нейросетевые модели, реализующие морфемный разбор словоформ русского языка. Их эффективность оценивалась одновременно по двум аспектам: точности морфемного разбора и затратам по времени работы и памяти (по производительности, вычисляемой в словах в секунду, и по объему памяти). Сверточная модель морфемного разбора словоформ показывает наилучшее достижимое качество морфемного разбора при достаточно высокой производительности, но требует заранее определенной части речи словоформ. Объединенная модель морфологического анализа дополнительно уточняет морфологические характеристики словоформ, в том числе часть речи. Предлагаемый комплекс подобных моделей позволяет достичь более высокой производительности морфемного разбора, но с некоторой потерей точности. Выбор модели для конкретной прикладной задачи зависит от особенностей последней. Реализованные модели встроены в открытый морфологический процессор русского языка[4].

Заметим, что производительность описанных моделей изучалась только с точки зрения архитектуры моделей машинного обучения. Дополнительное использование таких техник, как кэширование результатов анализа, квантование и удаление лишних весов, а также параллелизм может увеличить производительность на порядок.

Для обучения разработанных моделей были построены необходимые размеченные наборы данных (датасеты) со словоформами русского языка. В открытый доступ выложены как сами датасеты, так и реализованные модели морфологического анализа.[5]

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

[5] https://github.com/alesapin/XMorphy/tree/trying_tensorflow/scripts

Список литературы / References

[1] Большакова Е.И., Воронцов К.В. и др. Автоматическая обработка текстов на естественном языке и анализ данных: учебное пособие. Изд-во НИУ ВШЭ, 2017 г., 269 стр. / Bolshakova E.I., Vorontsov K.V. et al. Automatic processing of texts: handbook. HSE, 2017, 269 p. (in Russian)

[2] Ляшевская О.Н., Астафьева И. и др. Оценка методов автоматического анализа текста: морфологические парсеры русского языка. Труды международной конференции Диалог-2010, 2010, стр. 318-327 / Lyashevskaya O.N., Astafieva I. et al. Evaluation of automatic text analysis: morphological parsers for Russian. In Proc. of the International Conference Dialogue 2010, 2010, pp. 318-327 (in Russian).

[3] Harris Z.S. Morpheme boundaries within words: Report on a computer test. In Transformations and Discourse Analysis Papers. Formal Linguistics Series, Springer, 1970, pp. 68-77.

[4] Kanerva J., Ginter F. et al. Turku neural parser pipeline: An end-to-end system for the CoNLL 2018 shared task. In Proc. of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies, 2018, pp. 133-142.

[5] Anastasyev D.G. Exploring pretrained models for joint morpho-syntactic parsing of Russian. In Proc. of the International Conference Dialogue 2020, 2020, pp. 1-12.

[6] Sorokin A., Smurov I., Kirianov P. Tagging and parsing of multidomain collections. In Proc. of the International Conference Dialogue 2020, 2020, pp. 670-683.

[7] Lyashevskaya O.N., Shavrina T.O. et al. GRAMEVAL 2020 Shared Task: Russian Full Morphology and Universal Dependencies Parsing. In Proc. of the International Conference Dialogue 2020, 2020, pp. 553-569.

[8] Sorokin A., Kravtsova A. Deep convolutional networks for supervised morpheme segmentation of Russian language. Communications in Computer and Information Science, vol. 930, 2018, pp. 3-10.

[9] Bolshakova E., Sapin A. Comparing models of morpheme analysis for Russian words based on machine learning. In Proc. of the International Conference Dialogue 2019, 2019, pp. 104-113.

[10] Bolshakova E., Sapin A. Bi-LSTM Model for Morpheme Segmentation of Russian Words. Communications in Computer and Information Science, vol. 1119, 2019, pp. 151-160.

[11] Сокирко А.В. Морфологические модули на сайте www.aot.ru. Труды международной конференции Диалог-2004, 2004 г., стр. 559–564. / Sokirko A.V. Morphological components on www.aot.ru. In Proc. of the International Conference Dialogue 2004, 2004, pp. 559–564 (in Russian)

[12] Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages. Communications in Computer and Information Science, vol. 542, 2015, pp. 320-332.

[13] Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In Proc. of the International Conference on Machine Learning; Models, Technologies and Applications, 2003, pp. 273-280.

[14] Schmid H.: Probabilistic part-of-speech tagging using decision trees. In Proc. of the International Conference on New Methods in Language Processing, 1994, pp. 44-49.

[15] Straka M., Straková J., Hajic J. Prague at EPE 2017: The UDPipe system. In Proc. of the 2017 Shared Task on Extrinsic Parser Evaluation at the Fourth International Conference on Dependency Linguistics and the 15th International Conference on Parsing Technologies, 2017, pp. 65-74.

[16] Bojanowski P., Grave E. et al. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 2017, vol. 5, pp. 135-146.

[17] Peters M.E., Neumann M. et al. Deep contextualized word representations. In Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long Papers), 2018, pp. 2227–2237.

[18] Devlin J., Chang M.-W. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186.

[19] Kurimo M., Virpioja S. et al. Morpho challenge 2005-2010: Evaluations and results. In Proc. of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology, 2010, pp. 87-95.

[20] Virpioja S., Smit P. et al. Morfessor 2.0: Python implementation and extensions for Morfessor Baseline. Aalto University publication series science + technology, 2013, p. 38.

[21] Тихонов А.Н. Словообразовательный словарь русского языка. Русский язык, 1990 г., 864 стр. / Tikhonov A.N. Word Formation Dictionary of Russian language. Moscow, Russkiy yazyk, 1990, 864 p. (in Russian)

[22] OpenCorpora. URL: http://opencorpora.org/.
[23] Tensorflow – Large-Scale Machine Learning on Heterogeneous Systems. URL: https://www.tensorflow.org/.
[24] SynTagRus – Russian data from the SynTagRus corpus. URL: https://github.com/UniversalDependencies/UD_Russian-SynTagRus

Информация об авторах / Information about authors

Александр Сергеевич САПИН – аспирант кафедры алгоритмических языков факультета ВМиК. Сфера научных интересов: автоматическая обработка текстов на естественном языке, морфологический анализ и морфемный разбор слов в языках с богатой морфологией, применение машинного обучения для задач автоматической обработки текстов.

Alexander Sergeevich SAPIN is a post-graduate student of Algorithmic Languages Department, CMC Faculty. Research interests: natural language processing, morphological analysis and morpheme segmentation of words in natural languages with rich morphology, machine learning for NLP applications.