Автоматический морфемный разбор русских слов на основе решающих деревьев с применением бустинга

Сапин А.С., МГУ имени М.В. Ломоносова, факультет ВМК
alesapin@gmail.com

Аннотация

В статье представлена модель на базе машинного обучения для решения задачи морфемного разбора слов русского языка. Проводится обзор и экспериментальное сравнение с существующими решениями, при этом используются два словаря морфемного разбора русского языка отличающиеся по морфемному составу. Результаты экспериментов показывают, что представленная модель деревьев решений с применением бустинга превосходит другие модели для одного из рассматриваемых словарей, а для второго показывает сравнимое качество.

1 Введение

В задачах автоматической обработки текстов важнейшей составляющей является способ представления текста. Как правило, текст рассматривается как упорядоченный или неупорядоченный набор слов (лемм) или словоформ. Сами слова, в простейшем случае, могут быть представлены как бинарные вектора в пространстве слов, где размерностью вектора является количество всех слов в исследуемом тексте. Однако, такой подход является неэффективным из-за большой размерности пространства и вычислительной сложности. Для снижения размерности и более эффективного представления слов используются методы на базе машинного обучения.

Наиболее хорошо зарекомендовали себя методы на основе идей дистрибутивной семантики, такие как GloVe [Pennington, Socher, Manning, 2014] и Word2Vec [Mikolov, 2013]. Ещё одним методом получения векторных представлений слов является FastText [Joulin, 2016], который при обучении использует n-граммы букв текста, что дает улучшение качества в различных задачах компьютерной лингвистики. Возможным развитием этого метода является переход от использования n-грамм (не имеющих семантики) к морфемам (морфам), из которых состоит слово и которые являются минимальной смысловой единицей языка. Мы, как и в работах [Arefyev, Gratsianova, Popov, 2018; Sorokin, Kravtsova, 2018] предполагаем, что для морфологически богатых языков, каковым является русский, использование информации о морфемном составе слова может улучшить качество векторов слов, получаемых на основе дистрибутивных методов. Однако для обучения на базе этих методов необходима большая коллекция текстов, слова которой разбиты на морфы. Подобной коллекции нет пока ни для одного естественного языка. Для русского языка, например, есть лишь небольшие словари со словами разбитыми на морфы, что делает актуальной задачу автоматического морфемного разбора слов. Задачу получения морфемного разбора слова можно рассматривать как задачу классификации последовательностей букв. Её можно разделить на частную и более общую:

- Морфемная сегментация – разбиение слова на составляющие его морфы: *ВМЕН-Я-ТЬ*.
- Морфемная сегментация с классификацией – определение морфемного класса (типа) каждого сегмента (приставка, корень, суффикс и т.д.), что подразумевает нахождение границ между морфами: *ВМЕН:корень/Я:суффикс/ТЬ:окончание*

Для каждой из этих задач были предложены решения. Одно из первых решений задачи морфемной сегментации представлено в системе Morfessor [Creutz, Lagus, 2005]. Оно основано на машинном обучении без учителя с возможностью частичного использования уже размеченных данных, что обусловлено отсутствием больших словарей морфемного разбора для исследуемых языков.

Другая модель, показывающая лучшее качество решения задачи морфемной сегментации для русского языка, построена с использованием нейронных сетей типа encoder-decoder. Её описание содержится в статье [Arefyev, Gratsianova, Popov, 2018].

Также для русского языка было предложено решение на базе модели машинного обучения CRF, в рамках проекта CrossMorphy [Bolshakova, Sapin, 2017]. Обученная CRF-модель, при сегментации выделяет целиком группы морфем, например в слове *ЗВАННЫЙ* суффиксы *АН* и *Н* будут выделены вместе: *ЗВ:корень/АНН:суффиксы/ЫЙ:окончание*. Качество этой модели не уступает двум предыдущим реализациям.

В недавней работе [Sorokin, Kravtsova, 2018] была предложена более сложная модель на базе сверточных нейронных сетей, которая показала значительно лучшее качество решения задачи классификации с сегментацией.

В данной статье представлена новая модель морфемной сегментации с классификацией для слов русского языка, которая построена методом машинного обучения деревьев решений с градиентным бустингом (GBDT) [Dorogush, Ershov, Gulin, 2018]. Этот метод является более простым и интерпретируемым в сравнении с нейросетевыми методами, однако достаточно мощным для решения задач компьютерной лингвистики. В задачах классификации последовательностей важным является учёт влияния предыдущих и последующих элементов обучающей выборки на класс текущего элемента. Метод градиентного бустинга не работает с последовательностями, поэтому для учёта этого влияния в предложенной нами модели используется окно небольшого размера в предположении, что между буквами слова нет долгосрочных зависимостей.

Для обучения предложенной модели были использованы два словаря морфемного разбора русского языка, существенно различающихся по морфемному разбору слов, т.к. во многих случаях деление на морфы, например выделение суффиксов, выполняется неоднозначно. Первый словарь взят из проекта КроссЛексика [Большаков, 2013], который содержит около 23000 слов, второй словарь – это электронная версия словаря Тихонова [Тихонов, 2002], которая содержит около 90000 слов. Модель была обучена и протестирована на обоих словарях.

Тестирование показало, что в задаче сегментации наша модель достигает лучшего качества, чем известные решения, но в задаче сегментации с классификацией она лучше только для модели, обученной на словаре КроссЛексики, проигрывая модели на сверточных нейронных сетях, обученной на словаре Тихонова [Sorokin, Kravtsova, 2018].

2 Методы морфемной сегментации и классификации

Задача морфемной сегментации исследовалась достаточно давно. Первый метод её решения был представлен З. Харрисом в [Harris, 1970]. Его метод базируется на простой идее подсчета количества различных букв в словах текстового словаря, идущих после различных начальных частей слова и перед конечными частями слова. На Рисунке 1 приведен пример такого подсчета: в верхней строке находятся количества различных букв в словах словаря, идущих после начальных частей слова *интересный*, в нижней находятся количества перед конечными частями. Разбиение слова на морфемы происходит с помощью нахождения пиков (локальных максимумов) в каждом из рядов. В том месте, где обнаружен пик, и находится граница морфем слова. В рассмотренном примере такой пик выделяется между буквами *‘с’* и *‘н’*, что является корректной границей между морфемами. Также пик находится между буквам *‘р’* и *‘е’*, что не является действительной границей между морфемами. Данный метод был протестирован на небольшом словаре английского языка объёмом около 1000 слов и показал точность определения границ морфем около 61%.

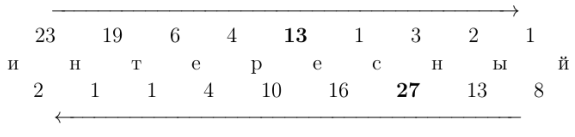

Рис. 1. Встречаемость различных букв после начальных и до конечных частей слова

Более известное решение задачи морфемной сегментации было реализовано в системе Morfessor [Creutz, Lagus, 2005]. Сегментация базируется на обучении с частичным привлечением учителя. На основании принципа минимальной длины описания (MDL) находится наилучшее морфемное разбиение для слов из заданного корпуса текстов на естественном языке. Так как данный метод допускает использование размеченных данных, сегментация может уточняться уже после обучения. Авторы обучали и тестировали свою модель для

английского и финского языка. Наилучшие результаты были показаны при обучении на неразмеченном корпусе в 200000 слов и дополнительном словаре с 10000 размеченными словами. Для финского языка лучшее значение F-меры по границам морфем составило 67.0% , а для турецкого – 70.7%.

В последнее время для задачи сегментации слов русского языка предложены и более сложные модели. В работе [Arefyev, Gratsianova, Popov, 2018] используется нейросетевая модель encoder-decoder, изначально созданная для задачи машинного перевода текстов. Данная модель превзошла результаты Morfessor, обученного для русского языка на корпусе lib.rus.ec [lib.rus.ec, 2018], на несколько процентов по точности.

Задача морфемной сегментации с классификацией исследовалась реже. Первые результаты для русского языка представлены в морфологическом процессоре CrossMorphy

[Bolshakova, Sapin, 2017], одна из функций которого – морфемный разбор слов путем классификации букв слов по основным типам морфов (приставка, корень, суффикс, окончание). По сути модель производит сегментацию с классификацией, но проводимая классификация достаточно груба, а сегментация не полная. Это, например, не позволяет увидеть границ между подряд идущими суффиксами и отличить постфикс *ся/сь* от окончания. Для классификации был использован метод обучения с учителем CRF [Lafferty, McCallum, Pereira, 2001]. Особенностью построения CRF-модели является использование в качестве признаков морфологических характеристик классифицируемых слов, а также статистических данных, получаемых методом Харриса. Модель была обучена и протестирована на двух словарях – словаре

КроссЛексики и словаре wikitionary [ru.wiktionary.org, 2016], состоящем в основном из данных словаря Тихонова. Наилучшая аккуратность (точность) классификации была показана на словаре КроссЛексики и составила 79.5%.

Модель более детальной морфемной сегментации с классификацией была предложена в [Sorokin, Kravtsova, 2018] и реализована на базе сверточных нейронных сетей (CNN). Модель протестирована и обучена на данных словаря Тихонова [Тихонов, 2002]. Для учета данных о последовательностях букв используется окно величиной 5 букв. Также применяется интересный метод сохранения информации о существующих морфемах словаря, называемый мемоизацией. Для каждой буквы особым образом кодируется информация о вхождении этой буквы в одну из морфем обучающей выборки и о том, находится ли эта буква на границе этой морфемы. Авторы пробовали добавлять LSTM-слои в сеть, однако это не дало улучшения качества модели. Наилучший результат модели составил 88.62% полностью верно разобранных слов.

3 Модель на базе деревьев с градиентным бустингом

В задаче морфемной сегментации с классификацией требуется не только определение класса каждой буквы, но и различение границ между морфемами. В русском языке проблема различения границ между подряд идущими морфемами особо сложна для суффиксов. Остальные виды морфем одного типа, в подавляющем большинстве случаев, не могут идти друг за другом. Выделяют 4 основных вида морфем – приставка (*P-PREF*), корень (*R-ROOT*), суффикс (*S-SUFF*) и окончание (*E-END*). При построении нашей модели сегментации с классификацией, также как и в модели [Sorokin, Kravtsova, 2018], мы вводим 3 дополнительных класса: соединительная гласная (*L-LINK*), дефис (*H-HYPH*) и постфикс (*PF-POSTFIX*). Для выделения границ суффиксов также добавлен класс *B-SUFF*(*BS*), обозначающий начало суффикса. Таким образом, модель классифицирует буквы слова на 8 классов, ниже показан пример классификации букв слова *учитель*:

у	*ч*	*и*	*т*	*е*	*л*	*ь*
B-ROOT	*ROOT*	*B-SUFF*	*B-SUFF*	*SUFF*	*SUFF*	*SUFF*

Для решения задачи классификации использовалась модель деревьев решений с градиентным бустингом [Friedman, 2002]. Используемые признаки делились на две категории: признаки, относящиеся к каждой отдельной букве, и признаки, относящиеся к слову.

Основным признаком была сама буква, представленная в формате one-hot encoding. Дополнительно использовались позиция буквы в слове, её гласность и частота встречаемости в словах словаря. Для отслеживания зависимостей между буквами применялось окно из соседних букв размером

в 10 (5 слева и 5 справа). Также в качестве признаков использовалась статистика по методу Харриса, показывающая количество различных букв в словаре после префикса, заканчивающегося текущей буквой, и перед суффиксом, начинающимся с текущей буквы.

Дополнительно к буквенным признакам были добавлены: длина слова и данные морфологического анализа, полученные морфологическим анализатором CrossMorphy [Bolshakova, Sapin, 2017]: учитывались часть речи, падеж, число, род, время и длина основы (из-за отсутствия информации о контексте сегментируемых слов для определения морфологических тегов в случае омонимии использовался простой статистический метод снятия морфологической омонимии, реализованный в CrossMorphy).

Например, буква *Р* из слова *НЕВЕРОЯТНЫЙ* будет иметь следующие значения признаков:

- Буква – ‘Р’
- гласность – *‘нет’*
- частота встречаемости – *0.4746*
- *буква[i-5]* – <*null*> (отсутствует)
- *буква[i-4] – Н*
- *буква[i-3] – Е*
- *буква[i-2] – В*
- *буква[i-1] – Е*
- *буква[i+1] – О*
- *буква[i+2] – Я*
- *буква[i+3] – Т*
- *буква[i+4] – Н*
- *буква[i+5] – Ы*
- статистика Харриса для начальной части слова *‘НЕВЕР’* – 7
- статистика Харриса для конечной части *‘РОЯТНЫЙ’* – 1
- часть речи – *ADJF* (полное прилагательное)
- падеж – accs (винительный)
- род – masc (мужской)
- число – sing (единственное)
- время – <*null*> (отсутствует)
- число букв в слове – 11
- длина основы слова – 9

Таким образом на вход модели подавалось 22 признака.

4 Эксперименты с моделью

4.1 Реализация модели

Для реализации модели была выбрана библиотека Catboost [Dorogush, Ershov, Gulin, 2018] для языка python, поскольку она показала лучшие результаты в сравнительном тестировании с другими библиотеками градиентного бустинга и не склонна к переобучению. Также в библиотечной реализации градиентного бустинга не требуется вручную кодировать в one-hot encoding категориальные признаки (такие как значения букв, части речи и т.п.), и в ней не требуется особым образом работать с числовыми признаками, что характерно для других реализаций методов градиентного бустинга. Библиотека предоставляет возможность обучения как на CPU, так и на GPU.

Для обучения и тестирования использовались данные из двух словарей морфемного разбора русского языка – словаря Тихонова объемом 96046 слов и словаря системы КроссЛексика объемом 23426 слова. Использование двух различных словарей обусловлено неоднозначностью правил морфемного разбора в русском языке, из-за чего словари составлены со значительными отличиями в трактовке суффиксов. Особенностью словаря Тихонова является то, что в нём содержится большое количество устаревших или узкоспециализированных слов, которые достаточно редко встречаются в текстах, например *остреквиться*, *розвязь*, *окулировать*. Словарь из проекта КроссЛексика содержит слова современного лексикона, однако разбор некоторых заимствованных слов значительно отличается от классического "школьного" разбора. Например, слова *дедукция*, *индукция* и *абдукция* имеют общий корень *дукц*, а *де-*, *ин-* и *аб-* соответственно выступают приставками. Стоит отметить, что в словаре проекта КроссЛексика отсутствуют многокоренные слова, а также слова с дефисом. Данные для обучения и тестирования делились в отношении 80:20, и обучающие выборки составили 76836 и 18740 слов соответственно.

4.2 Результаты экспериментов

В Таблице 1 приведены результаты экспериментов – в зависимости от словаря, использованного для обучения, глубины деревьев и типа вычислителя. Эксперименты проводились на двух различных системах: с CPU (Intel Xeon E5-2660v4, 256 GB RAM) и системе с графическим вычислителем (NVIDIA Tesla V100, 16 GB), однако модель, обученная на GPU показывала стабильно худшее качество. Дополнительно были проведены эксперименты с размером окна букв и количеством итераций, однако лучшие значения были достигнуты на окне размером 5 букв с каждой стороны и количеством итераций 10000 для словаря Тихонова и 5000 для словаря КроссЛексики, для оценки качества реализованной модели использовались 3 классических метрики – точность, полнота и F1-мера по границам морфем. Точность показывает отношение числа верно найденных границ морфем к общему числу найденных границ. Полнота показывает отношение числа верно найденных границ к общему числу границ. Дополнительно вычислялась аккуратность классификации, которая показывает отношение количества правильно определенных классов букв ко всем буквам. Точность по словам является интегральной метрикой и показывает отношение полностью корректно разобранных слов (и по границам и по классам букв) к количеству всех слов.

Табл. 1. Оценки построенной модели морфемного разбора

Словарь для обучения	Глубина	Вычислитель	Точность	Полнота	F1-мера	Аккуратность	Точность по словам
КроссЛексика	8	GPU	97.56	96.96	97.26	95.26	85.04
КроссЛексика	8	CPU	98.01	97.12	97.57	95.92	89.20
КроссЛексика	10	GPU	97.46	97.17	97.31	95.49	85.92
КроссЛексика	**10**	**CPU**	**98.63**	**98.42**	**98.53**	**97.55**	**92.04**
КроссЛексика	12	GPU	97.60	97.07	97.33	95.53	85.63
КроссЛексика	12	CPU	98.42	98.01	98.22	96.98	90.75
Сл. Тихонова	8	GPU	95.88	93.18	94.51	91.52	71.60
Сл. Тихонова	8	CPU	96.32	94.98	95.64	93.07	77.43
Сл. Тихонова	10	GPU	96.43	93.77	95.08	92.38	73.90
Сл.Тихонова	**10**	**CPU**	**97.90**	**95.56**	**96.71**	**94.95**	**81.62**
Сл.Тихонова	12	GPU	96.22	93.74	94.98	92.21	73.51
Сл.Тихонова	12	CPU	97.73	95.56	96.64	94.79	81.00

Из результатов видно, что морфемный разбор русского языка требует достаточно сложной модели. Качество постоянно улучшается в зависимости от глубины деревьев, однако на глубине 12 начинают проявляться эффекты переобучения, причем это не зависит от объема и сложности словаря, поэтому оптимальной глубиной модели является 10. Для обучения наилучшей модели на словаре КроссЛексики потребовалось около 5 часов на CPU и 2.5 минуты на GPU, а словаря Тихонова соответствующие результаты составили 23 часа и 4.3 минуты.

Так как модель деревьев решений является интерпретируемой, она позволяет увидеть вес признаков, который они вносят в результирующий ответ. Укажем важность признаков в процентах:

- Буква – 10.89

- Гласность – 4.18
- Позиция в слове – 1.77
- Частота встречаемости в словах – 4.30
- Буква[i − 5] – 3.5
- Буква[i − 4] – 3.55
- Буква[i − 3] – 5.74
- Буква[i − 2] – 9.19
- Буква[i − 1] – 11.41
- Буква[i + 1] – 11.79
- Буква[i + 2] – 7.32
- Буква[i + 3] – 4.77
- Буква[i + 4] – 3.35
- Буква[i + 5] – 1.67
- Число Харриса для начальной части – 4.19
- Число Харриса для конечной части – 3.12
- Часть речи – 2.44
- Падеж – 1.92
- Род – 1.27
- Число – 0.93
- Время – 0.52
- Число букв в слове – 1.01
- Длина основы слова – 1.13

Как мы видим, наибольший вес имеют сама классифицируемая буква, две следующих буквы и три предыдущих. Статистика Харриса также имела значительный вклад. Морфологические характеристики оказались менее важными, но в сумме имели вес 8.0%. В целом веса признаков оказывают достаточно ожидаемое влияние на результат.

Отметим, что до оценки качества к выходному результату модели дополнительно применялась процедура исправления некорректных последовательностей морфем на основе следующих очевидных правил: слово должно начинаться с приставки или корня, после приставки должен идти корень, после корня должен идти суффикс или окончание, после суффикса могут идти другой суффикс, кончание или постфикс.

Оказалось, что такой процедуре исправления подвергались около 230 слов из 4686 слов тестовой выборки словаря КроссЛексики и около 1000 из 19210 слов тестовой выборки для слова Тихонова, что составляет около 5%. Из такого сравнительно небольшого числа неправильных последовательностей классов морфем можно сделать вывод, что модель достаточно хорошо "понимает" принципы словообразования русского языка и основная часть ошибок приходится на неправильное определение границ морфем.

4.3 Анализ ошибок морфемного разбора

Анализ ошибок показал, что основное количество ошибок приходится на неверные определения границ между корнем и суффиксом, а также между суффиксами и окончаниями, например, для слова *ПЕЧЕЧКА ПЕЧЕЧ:корень/К:суффикс/А:окончание* вместо правильного *ПЕЧ:корень/ЕЧК:суффикс/А:окончание.*

Значительную часть ошибок также составляют ошибки неверного определения префикса. Чаще всего, он неверно определялся в словах, где его нет, например *ХО:префикс/ЗЯИН:корень*. Так как в русском языке количество приставок ограничено, данный вид ошибок может быть исправлен с помощью дополнительного правила корректирующего алгоритма.

Существенным недостатком исследуемой модели, который приводит к значительному проценту ошибок, является отсутствие классов для обозначения начал других морфем, кроме суффиксов. Поэтому редкие слова (многокоренные слова без соединительных гласных и со многими приставками), содержащие такие последовательности морфем, заведомо не могут быть разобраны верно. Однако, экспериментально было выяснено, что увеличение числа классов модели ведёт к снижению качества морфемного разбора при обучении обоих словарях.

5 Сравнение с другими моделями

Исследуемая модель решает общую задачу морфемной сегментации с классификацией на основные классы морфем. Результаты её работы можно сравнить как с моделями, решающими такую же задачу, так и с моделями, выполняющими только сегментацию. Для моделей сегментации наибольший интерес представляет F1-мера по границам морфем, а для общей задачи сегментации с классификацией интегральная метрика точности по словам.

Эксперименты показали, что наша модель на основе деревьев с градиентным бустингом превосходит модель из проекта Morfessor, обученную на корпусе lib.rus.ec, а также нейронную encoder-decoder модель, представленную в [Arefyev, Gratsianova,

Popov, 2018]. Несмотря на то, что метод CRF лучше подходит для анализа последовательностей, CRF-модель проигрывает модели градиентного бустинга. Результаты сравнения приведены в таблице 2.

Табл. 2. Значение F-меры по границам морфем

Модель	КроссЛексика	Слов.Тихонова
Градиентный бустинг	**98.53**	**95.08**
CRF от CrossMorphy	94.74	92.46
Morfessor	84.42	86.91
Ecoder-decoder	91.34	89.19

В задаче сегментации с классификацией, при обучении на словаре КроссЛексики, наша модель превосходит в точности по словам модель на базе сверточных нейронных сетей (CNN) [Sorokin, Kravtsova, 2018], однако при обучении на словаре Тихонова проигрывает ей – результаты показаны в таблице 3.

Табл. 3. Оценки точности по словам

Модель	КроссЛексика	Слов.Тихонова
CNN	91.58	**88.62**
Градиентный бустинг	**92.04**	81.62

Вероятной причиной проигрыша на словаре Тихонова может быть большое число слов, содержащих несколько подряд идущих корней (1038) и приставок (1197), другой причиной может бы недостаточность данных словаря КроссЛексики.

6 Заключение

Была разработана и исследована модель морфемной сегментации с классификацией, которая показывает хорошее качество и превосходит существующие модели на одном из обучающих словарей. Результаты работы такой модели предположительно могут быть полезны для построения векторного представления слов на основе таких методов, как FastText [Joulin, 2016]. Использование морфемы в качестве значащей единицы могло бы значительно улучшить качество таких методов. Проверка этой гипотезы является следующим шагом исследований.

Список литературы

Большаков И.А. 2013. КроссЛексика − Универсум связей между русскими словами. *Бизнес-информатика,* No 3, 2013.

Тихонов А. Н. 2002. Морфемно-орфографический словарь. М.: Издательство АСТ.

Arefyev N. V., Gratsianova T. Y., Popov K. P. 2018. Morphological Segmentation with Sequence to Sequence neural network. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018".*

Bolshakova E. I., Sapin A. S. 2018. A Morphological Processor for Russian with Extended Functionality. *International Conference on Analysis of Images, Social Networks and Texts, 2017.* LNCS, Springer.

Creutz M., Lagus K. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Helsinki : Helsinki University of Technology.

Dorogush A. V., Ershov V., Gulin A. 2018. CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363.*

Friedman J. H. Stochastic gradient boosting. 2002. *Computational Statistics & Data Analysis,* T. 38, №. 4.

Harris S. Zellig. 1970. Morpheme boundaries within words: Report on a computer test. *Transformations and Discourse Analysis Papers,* 73.

Joulin A. et al. Bag of tricks for efficient text classification. 2016. *arXiv preprint arXiv:1607.01759.*

Lafferty J., McCallum A., Pereira F. C. N. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Mikolov T. et al. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems.*

Pennington J., Socher R., Manning C. Glove: Global vectors for word representation. 2014. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).*

Sorokin A., Kravtsova A. 2018. Deep Convolutional Networks for Supervised Morpheme Segmentation of Russian Language. *Conference on Artificial Intelligence and Natural Language. CSIS, Springer.*

lib.ru.sec [Электронный ресурс]. URL: http://lib.rus.ec (дата обращения 10.12.2018).

ru.wiktionary.org [Электронный ресурс]. URL: https://ru.wiktionary.org/ (дата обращения 15.03.2016).