Создание метода проведения морфологического анализа несловарных слов в текстах на русском языке для повышения точности работы библиотек JMorfSdk

А. Н. Рыкунов, email: casi.05@mail.ru [1]

[1] Московский авиационный институт (национальный исследовательский университет)

Аннотация. *В работе рассматриваются методы анализа неизвестных слов, используемые в инструментах морфологического анализа, приводится алгоритм проведения морфологического анализа несловарных слов в библиотеке JMorfSdk, входящей в состав фреймворка TAWT. Описывается структура хранения необходимой информации для проведения такого анализа.*

Ключевые слова: *морфологический этап анализа, обработка естественного языка, компьютерная лингвистика, предсказание характеристик неизвестных слов.*

Введение

Количество текстовой информации постоянно растёт и необходимость автоматической её обработки приводит к созданию автоматизированных инструментов анализа текста, которые стали применяться во многих сферах деятельности людей, важными характеристиками используемых инструментов анализа текстов являются скорость и точность работы.

Сложность естественного языка привела к разделению его анализа на последовательные этапы, одним из которых является морфологический. Морфологический этап анализа текста отвечает за анализ отдельных слов, определение их частей речи и морфологических характеристик. Полученные данные в последующем используются на других этапах анализа, в частности, на этапе синтаксического анализа для построения синтаксической структуры текста [1].

Исторически в компьютерной лингвистике (КЛ) морфологический анализ выполнялся бессловарным методом. Преимуществом такого метода является возможность разбора абсолютно любого слова, но из-за большого количества исключений в естественном языке он является недостаточно точным.

В большинстве современных морфологических анализаторов используется метод анализа на основе словаря словоформ. Его

преимуществом является высокая точность, поскольку лексемы хранятся полностью. Но естественный язык постоянно развивается, в результате деятельности людей появляются новые сферы жизни и явления, и, следовательно, новые слова.

Постоянно пополнять словарь и поддерживать его в актуальном состоянии является трудной и даже не всегда выполнимой задачей, поэтому важной возможностью инструментов морфологического анализа является проведение морфологического анализа неизвестных слов. Таким образом современные системы совмещают словарный подход для слов, найденных в используемом инструментом словаре, и бессловарный подход для неизвестных слов.

1. Библиотека JMorfSdk в составе фреймворка TAWT

Фреймворк TAWT (Tools for Automated Work with Text) включает в себя набор программных инструментов лингвистического анализа текстов на русском языке [2]. Все инструменты используют общую схему подключения, стандартную для платформы Java: используется глобальный репозиторий бинарных зависимостей, исходный код которых, примеры и ссылки на артефакты находятся в общем доступе на Github [3].

Одним из модулей фреймворка является инструмент JMorfSdk (Java Morphological Software Development Kit), которые реализует морфологический этап анализа текста. Для разбора используется словарь проекта OpenCorpora [4], который основан на грамматическом словаре русского языка Зализняка, расширенном с помощью средств фреймворка TAWT, и включает в себя около 410 тысяч лексем и более 5 млн словоформ.

Алгоритмы анализа, реализованные в библиотеке JMorfSdk, имеют высокую производительность за счет использования хэш-таблиц вместе с использованием битовых операций и хранением самых необходимых характеристик в битовой шкале, что позволило получить константную сложность определения множества омоформ слова и их морфологических характеристик. Средняя скорость выполнения морфологического анализа текста составляет 0,9 миллионов слов/с. Отличительной особенностью инструмента является наличие режима генерации слов по заданным морфологическим характеристикам [5]. В статье, включающей в себя сравнительный анализ средств программных морфологической обработки [6], JMorfSdk стал лидером по скорости проведения анализа.

Однако, в библиотеке отсутствует возможность проведения морфологического анализа слов, не содержащихся в словаре. В сравнительном анализе работы морфологических анализаторов [6] у

большинства инструментов одним из недостатков был сравнительно небольшой размер словаря, что не позволяет провести полный морфологический анализ текста. Для повышения полноты и точности морфологического анализа был разработан алгоритм разбора несловарных слов, который позволит предсказывать части речи и характеристики неизвестных слов.

2. Методы морфологического анализа неизвестных слов в сторонних инструментах

Задача морфологического анализа несловарного слова сводится к определению характеристик, которыми может обладать анализируемое слово.

Некоторые слова могут заканчиваться, например, на «а», «о» или иметь нулевое окончание. Если производить анализ только по флексии, то будет получено огромное количество вариантов, поскольку такие флексии могут соответствовать разным частям речи. Для более точного предсказания характеристик помимо самой флексии также учитывается суффикс, стоящий перед ней. Под окончанием в рамках данной работы понимается часть слова, которая идёт после корня, включающая в себя суффиксы и флексию.

В русском языке присутствует несколько способов словообразования: приставочный, суффиксальный и приставочно-суффиксальный. Любое слово можно разбить на последовательность приставки, корня и окончания, где первое и последнее могут отсутствовать. Таким образом для определения характеристик также используются методы отсечения приставки и предсказания по окончанию.

В инструментах pymorphy2 [7] и АОТ [8] вначале анализа неизвестного слова происходит попытка отсечения приставки вместе с поиском по словарю. Особенностями методов отсечения приставок в этих морфологических анализаторах является то, что при неудачной попытке отсечения известных приставок перед переходом к анализу по окончанию слова происходит попытка отсечения части слова в начале, которая может являться неизвестной приставкой. Такой подход может привести к ошибочным результатам, если часть корня будет принята за приставку.

Следующим методом анализа неизвестного слова является предсказание по окончанию. Помимо уже рассмотренных инструментов такой подход реализован и в RussianMorphology [9], в котором он является единственным. Слова с одинаковым окончанием можно отнести к одной парадигме словообразования, из чего можно сделать вывод о морфологических характеристиках, как самого неизвестного

слова, так и всех словоформ лексемы. Работа данного метода заключается в определении окончания слова, в результате чего будет получена парадигма словообразования. Поскольку парадигма словообразования указывает на изменения в окончании, то остальные словоформы можно получить путём замены окончания одной словоформы на другую, с сохранением основной части слово нетронутой.

3. Разработанный метод морфологического разбора неизвестных слов

Были разработаны структуры хранения информации о приставках и окончаниях. Эти данные также, как и словарь представляют собой бинарный файл, что снижает количество занимаемой памяти, которое необходимо при запуске JMorfSdk Структуры хранения данных о приставках и окончаниях представлены на рисунках 1 и 2.

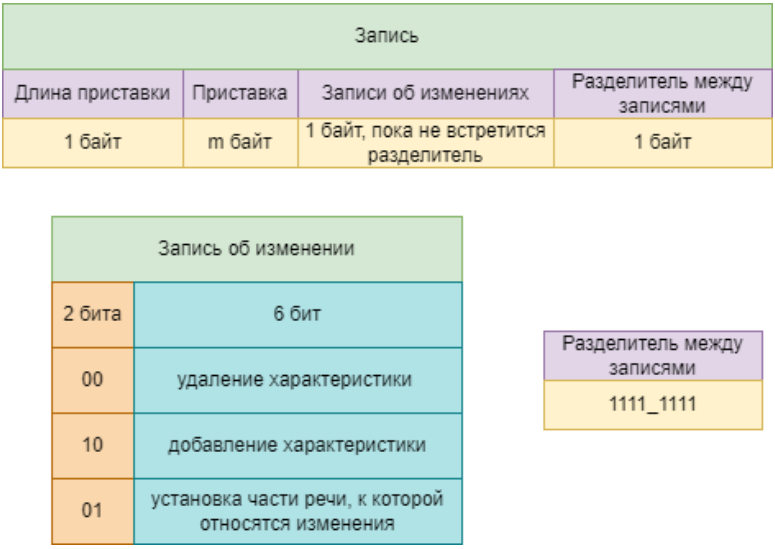

Рис. 1. Структура хранения данных о возможных приставках, а также их влиянии на морфологические характеристика

На основе проведённого анализа экспериментально было определено, что при приставочном методе словообразования происходит изменения только в характеристике «вид» у глаголов. При этом такой особенностью обладают только исконно-русские приставки. Приставки иноязычного происхождения влияют только на смысловую часть слова.

Окончание не принадлежит начальной форме						
Длина окончания	Окончание	Часть речи	Окончание начальной формы	Окончание	Морфологические характеристики	Морфологические характеристики начальной формы
1 байт	n байт	1 байт	1 байт	m байт	8 байт	8 байт

Окончание принадлежит начальной форме			
Длина окончания	Окончание	Часть речи	Морфологические характеристики начальной формы
1 байт	n байт	1 байт	8 байт

Рис. 2. Структура хранения данных об окончаниях, морфологических характеристиках, а также окончании и характеристиках начальной формы

Предложенная структура хранит как окончание и его характеристики, так и окончание начальной формы, на которое при запросе начальной формы будет произведена замена, и её характеристики. Хранение окончаний позволяет предсказать морфологические характеристики неизвестного слова, поскольку у слов с одинаковыми окончаниями обычно одна парадигма словоизменения.

Если при проведении морфологического анализа словоформа не была найдена в словаре, то сначала будет произведён поиск слов с отсечённой приставкой. В представленном алгоритме не происходит попытки анализа по неизвестным приставкам. Если метод отсечения приставок не дал результата, то происходит анализ по окончанию слов.

Окончание слова часто является указанием не только на определённую часть речи, но и на его характеристики. Например, суффикс «–ник–» указывает, что частью речи данного слова является имя существительное, а окончание будет влиять на число и падеж [10].

Иногда в анализируемом тексте могут встречаться слова, написанные через дефис. Хотя в используемом библиотекой словаре хранится большое количество слов, написанных через дефис, существует огромное количество вариантов, все из которых невозможно хранить в словаре. Поэтому важной частью анализа неизвестных слов является возможность разбора слов с дефисом. В разработанном методе они обрабатываются по специальным правилам. Если одна из частей является числом или одной из известных приставок, то анализ будет происходить по оставшейся части. В большинстве случаем первая часть слова является вспомогательной и меняет лишь смысл, в то время как

вторая часть определяет характеристики. В этом случае разбор достаточно провести по второй части.

4. Анализ результатов работы предложенного метода

Представленный метод анализа неизвестных слов использует только словарь приставок и окончаний, поэтому обладает преимуществом бессловарного подхода морфологического анализа. Бессловарный подход заключается в поиске наиболее длинного окончания и получение морфологических характеристик, соответствующих найденному окончанию. Следовательно, проведение морфологического анализа возможно для любого слова, для которого нашлось нужное окончание.

Попытка получения морфологических характеристик с помощью отсечения приставки перед переходом к анализу неизвестного слова с помощью метода предсказания по окончанию, позволяет ускорить процесс анализа, без потери в качестве разбора. Если к слову добавлена приставка, то оно будет иметь те же морфологические характеристики. За исключением некоторых исконно-русских приставок, которые могут влиять на вид глаголов, и приставки «по-», которая присутствует в наречиях. При её отсечении часть речи слова будет изменена на имя прилагательное.

Заключение

Проведённый ранее анализ работы инструментов [6] выявил ряд недостатков библиотеки JMorfSdk. Их исправление позволило значительно увеличить качество работы. Добавление в библиотеку возможности проведения морфологического анализа неизвестных слов также позволит повысить точность как получения морфологических характеристик определённой словоформы, так и снятия омонимии, которое в фреймворке реализовано на основе контекста. Точное получение характеристик неизвестного слова позволит повысить точность снятия омонимии во всём предложении.

Был предложен алгоритм морфологического анализа неизвестных слов вместе со структурой хранения информации для его проведения. Преимуществом алгоритма является возможность проводить морфологический анализ любого слова, окончание которого есть в словаре. Анализ происходит последовательно в два этапа – отсечение известных приставок и, если не были найдены морфологические характеристики, предсказание по окончанию.

Предложенные структуры данных и разработанный метод морфологического разбора неизвестных слов позволят наряду с

получением высокой точности морфологического анализа несловарных слов сохранить высокую скорость работы.

Литература

1. Mohbey K. K., Tiwari S. Preprocessing and morphological analysis in text mining // International Journal of Electronics Communication and Computer Engineering. - 2011. - Vol. 2. - № 2. - P. 1-7.
2. Politsyna E., Politsyn S., Porechny A. Solving practical tasks of computer linguistics using the created text processing framework [Электронный ресурс]: статья. – Дата обращения: 15.12.2021 – Режим доступа: https://iopscience.iop.org/article/10.1088/1742-6596/1902/1/012129
3. Официальная страница JMorfSdk. – Режим доступа – https://github.com/jalexpr/jmorfsdk. – (Дата обращения: 12.12.2021).
4. Официальный сайт OpenCorpora. – Режим доступа – http://opencorpora.org. – (Дата обращения: 12.12.2021).
5. Politsyna E. V. Development of the Cross-platform Library of Morphological Analysis of the Russian Language Text for Industrial Software / E. V. Politsyna, S. A. Politsyn, A. S. Porechny // CEE-SECR '18 Central and Eastern European Software Engineering Conference Russia Moscow. – ACM New York, NY, USA, 2018.
6. Рыкунов, А. Н. Исследование инструментов морфологического анализа текстов на русском языке для повышения точности алгоритмов обработки в библиотеке JMorfSdk / А. Н. Рыкунов, Е. В. Полицына, С. А. Полицын, А. С. Поречный // Информатика: проблемы, методы, технологии. – 2022. – С. 1204-1212.
7. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts. - 2015. - P. 320-332.
8. Официальный сайт AoT. – Режим доступа – http://aot.ru. – (Дата обращения: 12.12.2021).
9. Официальная страница RussianMorphology. – Режим доступа – https://github.com/AKuznetsov/russianmorphology. – (Дата обращения: 12.12.2021).
10. Морфемика и словообразование русского языка: учебно-методическое пособие для студентов, обучающихся по направлению подготовки 44.03.01 Педагогическое образование профиль «Начальное образование» / сост. Хертек Л. К. – Кызыл: Изд-во ТувГУ, 2018. – 131 с.