УДК 004.853

Н. И. Гурин, Я. А. Жук
Белорусский государственный технологический университет

МОРФОЛОГИЧЕСКИЙ АНАЛИЗ ТЕКСТА ДЛЯ ГЕНЕРАЦИИ БАЗЫ ЗНАНИЙ ДИАЛОГОВОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ

Статья посвящена описанию реализации алгоритма морфологического анализа текста, необходимого для автоматизации наполнения базы знаний диалоговой информационной системы. Алгоритм основан на гибридном методе морфологического анализа. Данный метод выработан путем анализа недостатков распространенных подходов к определению морфологических признаков слов на основе словарей и правил. Гибридный метод заключается в хранении в качестве правил псевдоокончаний, состоящих из произвольного числа последних символов в словах, при одинаковом изменении слов по формам и в хранении полных словоформ в случаях, когда формы слова абсолютно разные по написанию. Также описывается структура баз данных, используемых для определения морфологических признаков слов, разработанная на основе использующихся в языке способов образования форм слов. Алгоритм реализован на языке Python и осуществляет морфологический анализ в ходе построения базы знаний диалоговой информационной системы. Запросы на выборку из баз данных выполняются при помощи оператора REGEXP для сравнения анализируемого слова с псевдоокончаниями из баз данных, которые дополняются обозначением произвольного сочетания символов в качестве псевдоосновы. Дополнительным преимуществом алгоритма является кеширование результатов.

Ключевые слова: морфологический анализ, регулярные выражения, обработка естественного языка.

N. I. Gurin, Ya. A. Zhuk
Belarusian State Technological University

TEXT MORPHOLOGICAL ANALYSIS FOR THE DIALOG INFORMATION SYSTEM KNOWLEDGE BASE GENERATION

The article describes the implementation of the morphological text analysis algorithm needed for automation of the dialog information system knowledgebase filling. The algorithm is based on a hybrid method of morphological analysis. This method is developed by analyzing the disadvantages of common approaches to the word morphological characteristics definition based on dictionaries and rules. The hybrid method consists in storing rules in the form of pseudoendings including an optional number of word last characters when word forms has similar pseudobasis, and in storing full word forms when word forms are absolutely different in writing. The article also describes the database structure used to define the morphological characteristics of words, based on the methods of word forms formation used in the language. The algorithm is implemented in Python and performs the morphological analysis in the dialog information system knowledgebase building. The retrieval from the databases is performed using the REGEXP operator for searching matches of analyzed words with pseudoendings from databases, supplemented with notation of any combination of characters as pseudobasis. The additional benefit of the algorithm is the caching of results.

Key words: morphological analysis, regular expressions, natural language processing.

Введение. Для построения диалоговой информационной системы, способной отвечать на поставленные вопросы, необходимо обеспечить такую систему базой знаний, содержащей понятия предметной области и связи между ними. Полезность диалога с системой во многом зависит от объема базы знаний, однако наполнение базы знаний остается трудоемким процессом, выполняемым вручную. Автоматизировать данный процесс возможно путем автоматического извлечения знаний из текстов по заданной предметной области. Одним из начальных этапов семантического анализа исходного текста является морфологический анализ слов, т. е. определение части речи слова и его формы в предложении. Данная задача возникает как при разбиении сложных предложений на простые для различения причастных и деепричастных оборотов речи от перечислений, так и при семантическом анализе простых предложений в связи с тем, что элементом перехода, стоящим между двумя смысловыми единицами, с точки зрения актуального членения предложения является глагол [1].

В целом морфологический анализ слов исходного текста – важный этап для безошибочного

наполнения базы знаний и создания на ее основе семантической сети ключевых понятий диалоговой информационной системы.

Основная часть. Одним из направлений определения морфологических признаков слов является составление словарей, содержащих различные формы слов и их морфологические признаки. Недостатками данного метода считаются большая трудоемкость составления словаря, значительный объем словаря в памяти и неспособность к обработке слов, отсутствующих в словаре. Например, словарь проекта OpenCorpora составляет 1 344 568 слов и занимает 481,4 МБ [2].

Также существует направление определения морфологических признаков слов по правилам. Как отмечают разработчики универсальных анализаторов текстов, при наличии большого числа правил возникают конфликты между ними. Решением данной проблемы могут служить сортировка правил в определенном порядке либо составление деревьев решений.

Развитием методов на основе правил являются статистические методы, в которых для различных правил дается не один точный ответ, а набор вероятностей, и методы на основе грамматик, в которых правила объединяются в цепочки или деревья. Оба данных направления являются сложными в разработке и ресурсоемкими в применении [2].

Стоит отметить наличие различных способов реализации рассмотренных методов морфологического анализа. Основными отличиями данных способов являются форма хранения словаря или набора правил и средства реализации поиска морфологических признаков по форме слова. Наиболее распространенными приемами реализации считаются применение реляционных баз данных для хранения вместе с запросами на языке SQL для поиска и использование XML-файлов для хранения вместе со специализированными программными библиотеками для поиска.

Для разрабатываемого генератора семантических сетей на основе баз данных был применен гибридный подход, в котором в качестве основы выступает перечень правил в виде наборов из последних букв слов (псевдоокончаний или финалий [3]) в различных формах без привязки к определенным морфемам. Для разрешения конфликтов между правилами используется сортировка псевдоокончаний в порядке убывания их длины. Таким образом, при неправильном определении части речи или формы слова по существующим правилам можно добавить формы этого слова в набор правил целиком, как в словарь, чтобы его длина превысила длину использующихся правил. Такой подход позволяет сократить трудоемкость составления словаря и расход памяти на его хранение, а также избежать противоречий между правилами.

Наличие большинства морфологических признаков зависит от части речи, к которой принадлежит слово. Глаголы обладают спряжением, числом, временем, родом в прошедшем времени и лицом – в настоящем и будущем. Вспомогательные части речи, наречия и обозначенные цифрами числительные являются неизменяемыми частями речи. Остальные основные части речи (существительное, прилагательное, местоимение) изменяются по падежам и числу. Поэтому для хранения правил, описывающих перечисленные три группы слов, используются три таблицы реляционных баз данных, каждая из которых хранит наборы последних букв для каждой из обозначенных групп слов. В поля таблиц заносятся наборы последних символов для различных форм слов. Составленная структура баз данных представлена на рис. 1.

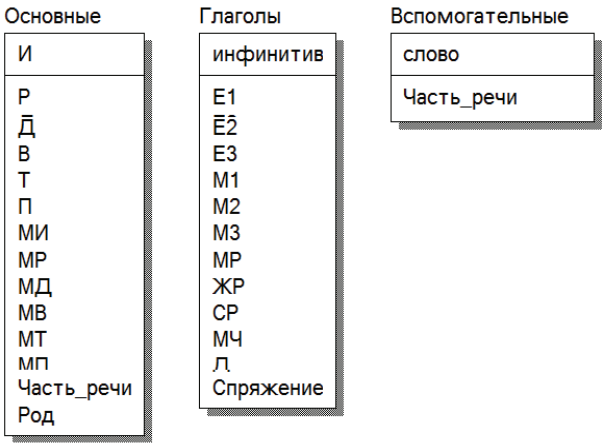

Рис. 1. Поля таблиц баз данных

Как видно из рис. 1, в таблице глаголов было создано 12 полей для псевдоокончаний форм слов (инфинитив, 2 числа по 3 лица, 3 рода и множественное число для прошедшего времени, деепричастие), в таблице несклоняемых частей речи – одно поле для единственной формы слова, в таблице основных частей речи – 12 полей (2 числа по 6 падежей). Кроме части речи, соответствующей финалиям в различных формах, в таблице хранятся другие морфологические признаки, такие как спряжение глаголов и род других основных частей речи.

Поскольку в таблице глаголов хранятся правила только для одной части речи, поле «Часть_речи» в данной таблице отсутствует в отличии от других таблиц, в которых хранятся слова различных частей речи. Стоит отметить,

что неизменяемые деепричастия вносятся отдельным полем в таблицу глаголов, а действительные и страдательные причастия записываются отдельно в таблицу основных частей речи, так как они изменяются по падежам и числу.

Для поиска морфологических признаков слова оно сравнивается при помощи регулярных выражений с псевдоокончаниями слов из таблиц баз данных. Данная операция реализуется при помощи соответствующих запросов на языке SQL, содержащих оператор сравнения строки с регулярным выражением REGEXP. Фрагмент листинга функции, выполняющей сравнение переданного в качестве аргумента слова с финалиями из базы данных глаголов с учетом возможности наличия постфикса «ся», приведен на рис. 2.

Следует отметить, что результаты запросов к трем таблицам затем сравниваются друг с другом для выявления наибольшего по длине. В то же время для неизменяемых слов выполняется дополнительный запрос на выборку точных совпадений слова с записями в базе данных для того, чтобы избежать некорректного распознавания коротких предлогов как наречий, оканчивающихся на сочетание символов, совпадающее с предлогом.

```
def compareStrWithVerbsEndings(s):
return "'" + s + "' REGEXP
CONCAT('.*',инфинитив,'$') OR '" + s + "'
REGEXP CONCAT('.*',E1,'$') OR '" + s + "'
REGEXP CONCAT('.*',E2,'$') OR '" + s + "'
REGEXP CONCAT('.*',E3,'$') OR '" + s + "'
REGEXP CONCAT('.*',M1,'$') OR '" + s + "'
REGEXP CONCAT('.*',M2,'$') OR '" + s + "'
REGEXP CONCAT('.*',M3,'$') OR '" + s + "'
REGEXP CONCAT('.*',Д,'$') OR '" + s + "'
REGEXP CONCAT('.*',инфинитив,'ся','$')
OR '" + s + "' REGEXP CONCAT('.*',E1,'ся','$')
OR '" + s + "' REGEXP CONCAT('.*',E2,'ся','$')
OR '" + s + "' REGEXP CONCAT('.*',E3,'ся','$')
OR '" + s + "' REGEXP CONCAT('.*',M1,'ся','$')
OR '" + s + "' REGEXP CONCAT('.*',M2,'ся','$')
OR '" + s + "' REGEXP CONCAT('.*',M3,'ся','$')
OR '" + s + "' REGEXP CONCAT('.*',Д,'ся','$')
OR '" + s + "' REGEXP
```

Рис. 2. Фрагмент листинга функции поиска глаголов

Кроме того, с целью оптимизации затрат вычислительных ресурсов на выполнение поиска результаты для каждого слова сохраняются в отдельном списке (кеше). Блок-схема программного модуля, реализующего функции определения части речи, к которой принадлежит слово, представлена на рис. 3.

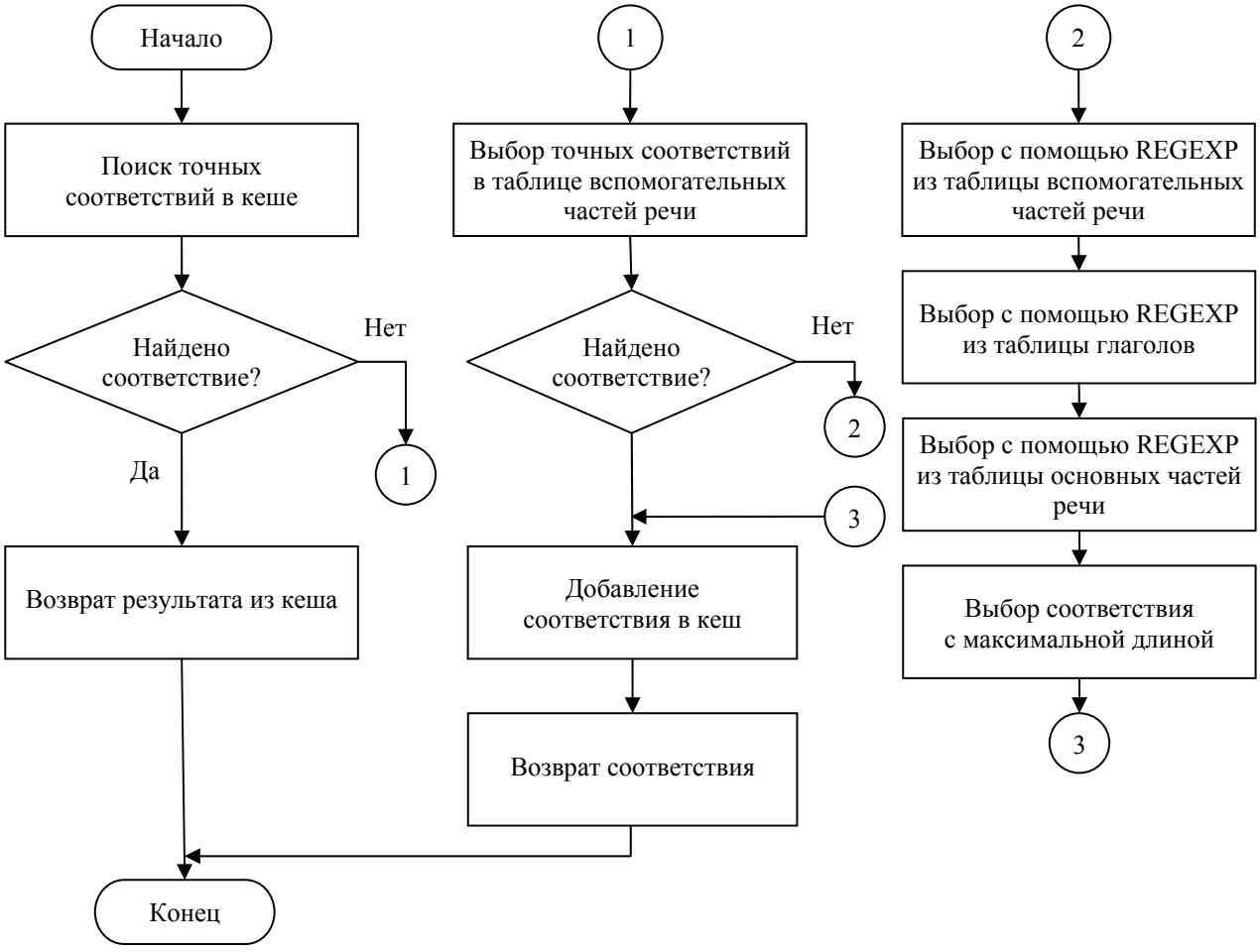

Рис. 3. Блок-схема алгоритма определения части речи слова при помощи регулярных выражений

В ходе апробации реализованной на языке Python функции морфологического анализа текста было обнаружено улучшение качества определения части речи слов технического текста по сравнению с предыдущей версией анализатора [4]: ряд слов в именительном и родительном падежах множественного числа были распознаны как существительные, хотя ранее по ошибке распознавались как наречия.

Заключение. Задача определения части речи является наиболее значимой среди задач морфологического анализа на этапе анализа текста для наполнения базы знаний. На этапах записи извлеченных знаний в семантическую сеть и синтеза предложений на основе извлеченных знаний возникают и другие задачи морфологического анализа, такие как определение формы слова, а также приведение слова в заданную форму. Разработанная функция морфологического анализа текста используется в составе генератора семантической сети для распознавания начала причастных и деепричастных оборотов, а также для автоматизации создания шаблонов предложений, выражающих различные типы семантических связей.

Литература

1. Лингвистический энциклопедический словарь / гл. ред. В. Н. Ярцева. М.: Сов. энциклопедия, 1990. 685 с.

2. Сегментация текста в проекте «Открытый корпус» / В. В. Бочаров [и др.] // Компьютерная лингвистика и интеллектуальные технологии: материалы Междунар. конф., Бекасово, 30 мая – 3 июня 2012 г. В 2 т. Т. 1: Основная программа конференции / РГГУ. М., 2012. С. 51–60.

3. Большаков И. А., Большакова Е. И. Автоматический морфоклассификатор русских именных групп // Компьютерная лингвистика и интеллектуальные технологии: материалы Междунар. конф., Бекасово, 30 мая – 3 июня 2012 г. В 2 т. Т. 1: Основная программа конференции / РГГУ. М., 2012. С. 81–92.

4. Гурин Н. И., Жук Я. А. Генератор семантической сети информационной системы в таблицу реляционной базы данных // Труды БГТУ. 2015. № 6: Физ.-мат. науки и информатика. С. 181–185.

References

1. *Lingvisticheskiy entsiklopedicheskiy slovar'* [Linguistic encyclopedic dictionary]. Moscow, Sovetskaya entsiklopedyia Publ., 1990. 685 p.

2. Bocharov V. V., Alekseeva S. V., Granovskiy D. V., Ostapuk N. A., Stepanova M. E., Surikov A. V. [Text segmentation in "OpenCorpora"]. *Materialy mezhdunarodnoy konferentsii (Komp'yuternaya lingvistika i intellektual'nyye tekhnologii)* [Materials of the International Conference (Computer linguistics and artificial intelligence)]. Bekasovo, 2012, pp. 51–60 (In Russian).

3. Bol'shakov I. A., Bol'shakova E. I. [Automated morphoclassifier of russian nominal groups]. *Materialy mezhdunarodnoy konferentsii (Komp'yuternaya lingvistika i intellektual'nyye tekhnologii)* [Materials of the International Conference (Computer linguistics and artificial intelligence)]. Bekasovo, 2012, pp. 81–92 (In Russian).

4. Gurin N. I., Zhuk Ya. A. The information system semantic network generator to a relational database table generator. *Trudy BGTU* [Proceedings of BSTU], 2015, no. 6: Physical-mathematical sciences and informatics, pp. 181–185 (In Russian).

Информация об авторах

Гурин Николай Иванович − кандидат физико-математических наук, доцент кафедры информационных систем и технологий. Белорусский государственный технологический университет (220006, г. Минск, ул. Свердлова, 13а, Республика Беларусь). E-mail: ngourine@mail.ru

Жук Ярослав Александрович − аспирант. Белорусский государственный технологический университет (220006, г. Минск, ул. Свердлова, 13а, Республика Беларусь). E-mail: zhuk@belstu.by

Information about the authors

Gurin Nikolay Ivanovich − PhD (Physics and Mathematics), Assistant Professor, the Department of Information Systems and Technologies. Belarusian State Technological University (13a, Sverdlova str., 220006, Minsk, Republic of Belarus). E-mail: ngourine@mail.ru

Zhuk Yaroslav Aleksandrovich − PhD student. Belarusian State Technological University (13a, Sverdlova str., 220006, Minsk, Republic of Belarus). E-mail: zhuk@belstu.by

Поступила 03.03.2016