АНАЛИЗ ПРОГРАММНЫХ СРЕДСТВ МОРФОЛОГИЧЕСКОГО АНАЛИЗА

В.Д. Шульман[1], *магистр*
О.Е. Максименко[2], *магистр*
П.Д. Волхонцева[1], *магистр*
[1]**Московский государственный технический университет им. Н.Э. Баумана**
[2]**Национальный ядерный исследовательский университет Московский инженерно-физический институт**
(Россия, г. Москва)

DOI:10.24412/2500-1000-2022-3-2-166-170

Аннотация. *Статья посвящена анализу трех программных инструментов, применяемых для морфологического анализа текста – pymorphy2, myStem и Stanza. Дано определение морфологическому анализу и подчеркнута актуальность его применения при взаимодействии человека с ЭВМ. Приведены характеристики и возможности каждого из них. Сделаны выводы о целесообразности использования их при работе с разными текстовыми данными.*

Ключевые слова: *морфологический анализатор, обработка текста, словарь, грамматические характеристики, инфинитивная форма, аннотация, корпус текстов.*

Задача морфологической обработки текстов на естественных языках является актуальной для разных сфер жизни человечества, например голосовых помощниках или автоматических переводах текста. Алгоритмы такой обработки были реализованы и развиты после появления ЭВМ. Им можно дать общее название – Автоматическая обработка тестов (АОТ). Морфологический анализ можно рассматривать как один из процессов, составляющих данную обработку [1]. Этот процесс позволяет, к примеру, определять грамматические признаки неизвестных слов и получать их основы, что может быть использовано в рамках взаимодействия человека и ЭВМ.

Понятие морфологического анализатора. Морфологический анализ текста – это процесс определения грамматического значения словоформ и выделения их основ или по-другому – лемматизация [2].

Программные средства морфологического анализа представлены несколькими решениями, способными работать с разными языками программирования, мы же будем рассматривать те, которые имеют свою реализацию на Python.

MyStem – морфологический анализатор, разработанный компанией Яндекс. Первая версия была создана в 90-х годах, однако не имела большой популярности и не находилась в открытом доступе. Стоит отметить, что первая версия предполагала использование словаря небольшого размера, опираясь в основном на методы бессловарной морфологии, в то время как текущие реализации базируются на классическом подходе словарной морфологии. В частности, используется словарь Зализняка [3], в котором морфологические гипотезы о новых словах формируются при помощи префикса деревьев. В настоящий момент MyStem версии 3.0 предоставляет все функции полного морфологического анализа, однако не имеет функции синтеза. Данная версия является наиболее стабильной и доступной для скачивания в бинарном виде. Существует также пакет pymystem3, позволяющий работать с Mystem в Python.

Существуют также библиотеки, поддерживающие мультиязычность в своей работе. Примером подобной является Stanza. Stanza – это пакет анализа естественного языка. Он содержит инструменты, которые можно использовать в конвейере для преобразования строки, содержащей текст на человеческом языке, в списки предложений и слов, для создания базовых форм этих слов, их частей речи и морфологических признаков, для синтак-

сического анализа зависимостей структуры и распознавания именованных объектов [4]. Инструментарий разработан таким образом, чтобы он мог использоваться при работе с более чем 70 языками, используя формализм универсальных зависимостей.

Инструмент Stanza построен с использованием высокоточных компонентов нейронной сети, которые также обеспечивают эффективное обучение и оценку с использованием ваших собственных аннотированных данных. Модули построены поверх библиотеки Pitch. Можно получить гораздо более высокую производительность, если запустить программное обеспечение на компьютере с поддержкой GPU.

PyMorphy2 написан на языке Python (работает под 2.7 и 3.5+). Он умеет:

1. приводить слово к нормальной форме (например, “люди -> человек”, или “гулял -> гулять”).

2. ставить слово в нужную форму. Например, ставить слово во множественное число, менять падеж слова и т.д.

3. возвращать грамматическую информацию о слове (число, род, падеж, часть речи и т.д.).

При работе используется словарь OpenCorpora; для незнакомых слов строятся гипотезы.

OpenCorpora («Открытый корпус») – краудсорсинговый проект создания морфологически, синтаксически и семантически размеченного корпуса текстов на русском языке, в полном объёме доступном для исследователей [5]. Проект существует с 2009 года и содержит свободные тексты, распространяющиеся на условиях лицензии CC-BY-SA.

Отображение некоторых частей речи представлено в таблице 1.

Таблица 1. Отображение частей речи в анализаторах

Часть речи	MyStem	pymorphy2	Stanza
Глагол	V	V	VERB
Существительное	S, SPRO	NOUN, NPRO, LTN	NOUN
Прилагательное	A, A-NUM, A-PRO	ADJF, ADJS, COMP	ADJECTIVE
Наречие	ADV, ADVPRO	ADVB, PRED	ADVERB
Числительное	NUM	NUMB, NUMR, ROMN	NUM

Получение информации по текстовым данным. Для демонстрации возможностей обозреваемых анализаторов был выбран язык программирования Python и соответствующие библиотеки. Для демонстрации функционала в качестве входных данных подается предложение «мама меня любит». Ниже приведен пример анализа предложения от MyStem.

```
[{'analysis': [{'lex': 'мама', 'wt': 1, 'gr': 'S,жен,од=им,ед'}],
'text': 'мама'}, {'text': ' '}, {'analysis': [{'lex': 'я', 'wt':
0.9999549915, 'gr': 'SPRO,ед,1-л=(вин|род)'}], 'text': 'меня'},
{'text': ' '}, {'analysis': [{'lex': 'любить', 'wt': 1, 'gr':
'V,несов,пе=непрош,ед,изъяв,3-л'}], 'text': 'любит'}, {'text':
'\n'}]

S,жен,од=им,ед

SPRO,ед,1-л=(вин|род)

V,несов,пе=непрош,ед,изъяв,3-л

мама я любить
```

Рис. 1. Фрагмент кода с морфологической информацией, выдаваемой MyStem

В ответе приведены следующие поля:
- lex – лексема, отдельное слово;
- wt – вес;
- gr – грамматическая информация;
- text – исходное слово из текста.

Грамматическая информация содержит:
- код части речи;
- число;
- род;
- лицо;
- падеж.

Рассмотрим морфологический разбор от Stanza. Из всех рассматриваемых анализаторов этот – единственный, что поддерживает мультиязычность, выбор языка требуется дополнительно прописать в коде.


```
{
"id": 1,"text": "Мама","lemma": "мама","upos": "NOUN",
"feats": "Animacy=Anim|Case=Nom|Gender=Fem|Number=Sing",
"head": 3,"deprel": "nsubj","start_char": 0,"end_char":
4,"ner": "O"
},
{"id": 2,"text": "меня","lemma": "я","upos":
"PRON","feats": "Case=Acc|Number=Sing|Person=1","head":
3,"deprel": "obj","start_char": 5,"end_char": 9,"ner":
"O"
},
{"id": 3,"text": "любит","lemma": "любить","upos":
"VERB","feats":
"Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|
VerbForm=Fin|
Voice=Act","head":0,"deprel":"root","start_char": 10,
"end_char": 15, "ner": "O"
}
```


Рис. 2. Ответ от анализатора Stanza

Ответ содержит поля:
- id – позиция в тексте;
- text – исходное слово;
- lemma – лемма, слово в инфинитивной форме;
- upos – часть речи;
- feats – лексические и грамматические свойства слов;
- start_char – позиция первого символа слова в исходном тексте;
- end_char – позиция последнего символа слова в исходном тексте.

Грамматические свойства кодируются с помощью Universal Dependencies. Это основа для последовательного аннотирования грамматики (частей речи, морфологических особенностей и синтаксических зависимостей) в разных человеческих языках [6]. Основными приложениями являются автоматизированная обработка текста в области обработки естественного языка (NLP) и исследования синтаксиса и грамматики естественного языка, особенно в рамках лингвистической типологии. Основная цель проекта – добиться межъязыковой согласованности аннотаций, в то же

время допуская при необходимости расширения, зависящие от конкретного языка [7].

[Parse(word='мама', tag=OpencorporaTag('NOUN, anim, femn sing, nomn'), normal_form='мама', score=1.0, methods_stack=((DictionaryAnalyzer(), 'мама', 1988, 0),))]

[Parse(word='меня', tag=OpencorporaTag('NPRO,1per sing, accs'), normal_form='я', score=0.536184, methods_stack=((DictionaryAnalyzer(), 'меня', 3246, 3),)), Parse(word='меня', tag=OpencorporaTag('NPRO,1per sing, gent'), normal_form='я', score=0.463815, methods_stack=((DictionaryAnalyzer(), 'меня', 3246, 1),))]

[Parse(word='любит', tag=OpencorporaTag('VERB, impf, tran sing, 3per, pres, indc'), normal_form='любить', score=1.0, methods_stack=((DictionaryAnalyzer(), 'любит', 1967, 5),))]

Рис. 3. Информация от анализатора PyMorphy2

Рассмотрим синтаксис ответа PyMorphy2.

Структура ответа состоит из поля:

- word – исходное слово;
- tag – грамматические характеристики;
- normal_form – начальная форма слова;
- score – это оценка P(tag|word), оценка вероятности того, что данный разбор правильный.

Заключение. В рамках данной статьи был сделан обзор трех морфологических анализаторов, даны их программные характеристики, обозначена актуальность их применения в процессе автоматической обработки текстов. Для рассмотрения был взяты отечественные и иностранный программный продукт, способный работать не только с русским языком. Каждый из них предоставляет основной набор грамматической и лингвистической информации о слове, однако есть различия в дополнительных параметрах.

Библиографический список

1. Реформатский А.А. Введение в языковедение. – М.: Аспект Пресс, 2004. – 536 с.

2. Апресян Ю.Д., Богуславский И.М., Йомдин Л.Л. и др. Лингвистический процессор для сложных информационных систем. – М.: Наука, 1992. – 256 с.

3. Зализняк А.А. Грамматический словарь русского языка. – М., Русский язык, 1980. – 880 с.

4. Кожина М.Н. Речеведение и функциональная стилистика: вопросы теории. – Пермь, 2002. – 475 с.

5. Компьютерная семантика русского языка. – СПб.: Изд-во СпбГУ. – 20 с.

6. Морфологический анализатор текста на русском языке mystem // Компания Яндекс [сайт]. – 2003–2013. – [Электронный ресурс]. – Режим доступа: http://company.yandex.ru/technologies/mystem/ (дата обращения - 10.03.2022)

7. Пруцков А. В., Розанов А. К. Методы морфологической обработки текстов // Прикаспийский журнал: управление и высокие технологии. – 2014. – №3 (27). – С. 119-133.

ANALYSIS OF MORPHOLOGICAL ANALYSIS SOFTWARE TOOLS

V.D. Shulman[1], *Master*
O.E. Maksimenko[2], *Master*
P.D. Volkhontseva[1], *Master*
[1]Bauman Moscow State Technical University
[2]National Research Nuclear University Moscow Engineering Physics Institute
(Russia, Moscow)

Abstract. *The article is devoted to the analysis of three software tools used for morphological analysis of text – pymorphy2, myStem and Stanza. The definition of morphological analysis is given and the relevance of its application in human interaction with a computer is emphasized. The characteristics and capabilities of each of them are given. Conclusions are drawn about the expediency of using them when working with different text data.*

Keywords: *morphological analyzer, text processing, dictionary, grammatical characteristics, infinitive form, abstract, corpus of texts.*