УДК 378.14

# МЕТОДИКА ПРЕПОДАВАНИЯ ОСНОВ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

**Самохвалов А.Э.**

*ФГБОУ ВО «Московский государственный технический университет имени Н.Э. Баумана», Москва, e-mail: samox@bmstu.ru*

Целью исследования является составление методики преподавания основ обработки естественного языка, которая должна соответствовать сформулированной в Национальной стратегии развития искусственного интеллекта в Российской Федерации на период до 2030 г. стратегической задаче – повышению уровня обеспечения отечественного рынка технологий искусственного интеллекта квалифицированными кадрами. Образовательный модуль ориентирован на обучение студентов учебных заведений среднего специального и высшего образования. В работе применены методы сбора, анализа, обобщения информации о программных способах обработки естественного языка, нормативно-правовых документах в сфере образования, методиках проведения теоретических и практических занятий по компьютерной лингвистике и анализу текстов. Составлены тематические занятия, каждому из которых соответствуют универсальные и общепрофессиональные компетенции, перечисленные в Федеральном государственном образовательном стандарте высшего образования. Учебная программа составлена по принципу логической последовательности тем: изучение документации, инсталляция программного обеспечения, исследование свойств и методов морфологического анализатора, решение прикладных задач, практическая работа, отчет. Предложенный автором курс занятий прошел пилотную апробацию в Московском государственном техническом университете им. Н.Э. Баумана при подготовке бакалавров по направлениям 09.03.01 «Информатика и вычислительная техника» и 09.03.03 «Прикладная информатика» (профиль «Информационная аналитика»).

**Ключевые слова: искусственный интеллект, обработка естественного языка, образование, информатика, Pymorphy2**

# METHODS OF TEACHING THE BASICS OF NATURAL LANGUAGE PROCESSING

**Samokhvalov A.E.**

*Bauman Moscow State Technical University, Moscow, e-mail: samox@bmstu.ru*

The purpose of the study is to develop a methodology for teaching the basics of natural language processing, which should correspond to the strategic objective formulated in the National Strategy for the Development of Artificial Intelligence in the Russian Federation for the period up to 2030 – to increase the level of provision of qualified personnel to the domestic artificial intelligence technology market. The educational module is aimed at teaching students of secondary specialized and higher education institutions. The paper uses methods for collecting, analyzing, and summarizing information about software methods of natural language processing, regulatory documents in the field of education, and methods for conducting theoretical and practical classes in computational linguistics and text analysis. Thematic classes have been compiled, each of which corresponds to universal and general professional competencies listed in the Federal State Educational Standard of Higher Education. The curriculum is based on the principle of a logical sequence of topics: studying documentation, installing software, studying the properties and methods of a morphological analyzer, solving applied problems, practical work, and a report. The course of classes proposed by the author was piloted at the Bauman Moscow State Technical University for bachelor's degree programs in the fields of 09.03.01 "Informatics and Computer Engineering" and 09.03.03 "Applied Informatics" (Information Analytics profile).

**Keywords: artificial intelligence, natural language processing, education, computer science, Pymorphy2**

## Введение

В Национальной стратегии развития искусственного интеллекта (ИИ) в Российской Федерации на период до 2030 г. поставлена стратегическая задача – повысить уровень обеспечения отечественного рынка технологий ИИ квалифицированными кадрами. Для ее решения необходимы «разработка и внедрение образовательных модулей в рамках образовательных программ всех уровней образования, программ повышения квалификации и профессиональной переподготовки для получения гражданами знаний, приобретения ими компетенций и навыков в области математики, программирования, анализа данных, машинного обучения, способствующих развитию искусственного интеллекта» [1].

В современном информационном обществе наблюдается широкое применение ИИ в компьютерной лингвистике и обработке естественного языка (ОЕЯ), где одним из основных направлений применения является исследование текста: извлечение ключевых слов и понятий, кластеризация, обобщение и семантический анализ [2].

**Цель исследования** – разработать методику преподавания основ обработки естественного языка студентам образовательных организаций среднего специального и высшего образования.

**Материалы и методы исследования**

В работе исследованы нормативно-правовые документы, регулирующие образовательную деятельность, применены методы сбора, анализа, обобщения литературы, посвященной технологиям ОЕЯ; возможностям языка программирования Python и библиотек ИИ; методикам преподавания, проведения теоретических и практических занятий по изучению ИИ.

**Результаты исследования и их обсуждение**

Основой современных технологий ОЕЯ служат морфологические анализаторы текста [3], с помощью которых начинается обработка. Широкое распространение получил анализатор текста на русском языке Pymorphy2 [4].

Автор предлагает разделить процесс изучения возможностей морфологического анализатора текста Pymorphy2 на следующие этапы (темы занятий): изучение документации, инсталляция программного обеспечения, исследование свойств и методов класса MorphAnalyzer, применение морфологического анализатора Pymorphy2 при решении прикладных задач, практическая работа на заданную тему, отчет о НИР. Результатом представленного курса обучения является приобретение студентом универсальных и общепрофессиональных компетенций, перечисленных в Федеральном государственном образовательном стандарте высшего образования по направлению подготовки 09.03.01 «Информатика и вычислительная техника» [5] (таблица).

На занятии «Изучение документации» студенты знакомятся с официальной документацией по морфологическому анализатору Pymorphy2 на сайте https://Pymorphy2.readthedocs.io в следующей последовательности: терминология, обозначения для граммем на русском языке (часть речи, падеж, число, род, нестандартные граммемы), установка, морфологический анализ, работа с тегами, кириллические названия тегов, склонение слов, постановка слов в начальную форму, согласование слов с числительными, выбор правильного разбора [6].

Для программирования прикладных задач обучающимся необходимо прочитать документацию по библиотеке Natural Language Toolkit (NLTK) на официальном сайте https://www.nltk.org и по библиотеке Wordcloud на официальном сайте https://pypi.org/project/wordcloud.

Занятие «Инсталляция программного обеспечения» посвящено установке библиотеки Pymorphy2 для языка программирования Python и электронных словарей Pymorphy2-dicts-ru для обработки русского языка. Варианты инсталляции: для операционной системы MS Windows (согласно официальной документации разработчиков), для отечественной операционной системы Alt Linux (пакет python3-модуль-Pymorphy2).

Аналогично выполняется установка библиотек NLTK и Wordcloud.

Темы занятий и компетенции

| Тема занятия | Компетенции ФГОС |
|---|---|
| Изучение документации | УК-1. Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач |
| Инсталляция программного обеспечения | ОПК-5. Способен инсталлировать программное и аппаратное обеспечение для информационных и автоматизированных систем |
| Исследование свойств и методов класса MorphAnalyzer | ОПК-1. Способен применять естественнонаучные и общеинженерные знания, методы математического анализа и моделирования, теоретического и экспериментального исследования в профессиональной деятельности.<br>ОПК-2. Способен понимать принципы работы современных информационных технологий и программных средств, в том числе отечественного производства, и использовать их при решении задач профессиональной деятельности |
| Применение морфологического анализатора Pymorphy2 при решении прикладных задач | ОПК-9. Способен осваивать методики использования программных средств для решения практических задач |
| Практическая работа на заданную тему | ОПК-8. Способен разрабатывать алгоритмы и программы, пригодные для практического применения |
| Отчет о НИР | ОПК-4. Способен участвовать в разработке стандартов, норм и правил, а также технической документации, связанной с профессиональной деятельностью |

И сточник: составлено автором на основании Приказа Министерства образования и науки РФ от 19 сентября 2017 г. № 929 [5].

На занятии «Исследование свойств и методов класса MorphAnalyzer» студентам предлагается создать программу (ру-файл) в интегрированной среде программирования IDLE (включена в стандартный пакет «Альт Образование»), в которой с помощью следующей команды подключается библиотека морфологического анализатора текста:

import Pymorphy2

Создается объект класса морфологического анализатора MorphAnalyzer:

morph = Pymorphy2.MorphAnalyzer()

Выполняется метод Parse для получения информации по анализу слова, введенного как входной параметр, например: morph.parse(‘аналитические’). Результат выполнения метода:

[Parse(word=’аналитические’, tag=OpencorporaTag(‘ADJF plur,nomn’), normal_form=’аналитический’, score=0.5, methods_stack = ((DictionaryAnalyzer(), ‘аналитические’, 16, 20),)),

Parse(word=’аналитические’, tag=OpencorporaTag(‘ADJF inan,plur,accs’), normal_form=’аналитический’, score=0.5, methods_stack = ((DictionaryAnalyzer(), ‘аналитические’, 16, 24),))]

С помощью изученной на первом занятии документации обучающемуся предлагается расшифровать полученные тэги. Для приведенного выше примера получится следующая расшифровка:

1) слово «аналитические»; граммемы – имя прилагательное (полное), множественное число, именительный падеж; нормальная форма «аналитический»;

2) слово «аналитические»; граммемы – имя прилагательное (полное), неодушевленное, множественное число, винительный падеж; нормальная форма «аналитический».

С помощью метода lat2cyr студент выполняет перевод граммемы с английского языка на русский язык, например: morph.lat2cyr(‘ADJF plur,nomn’). Результат выполнения метода: ‘ПРИЛ мн,им’.

С помощью метода inflect выполнится склонение слова, например, в родительном падеже:

w = morph.parse(‘аналитические’)[0]

w.inflect({‘gent’})

Результат выполнения метода:

Parse(word=’аналитических’, tag=OpencorporaTag(‘ADJF plur,gent’), normal_form=’аналитический’, score=1.0, methods_stack= ((DictionaryAnalyzer(), ‘аналитических’, 16, 21),))

С помощью свойства lexeme студент выводит на экран лексему (все возможные формы слова) и расшифровывает полученные тэги.

С помощью свойства normal_form можно получить нормальную форму слова (именительный падеж, единственное число): w.normal_form. Результат выполнения кода: ‘аналитический’.

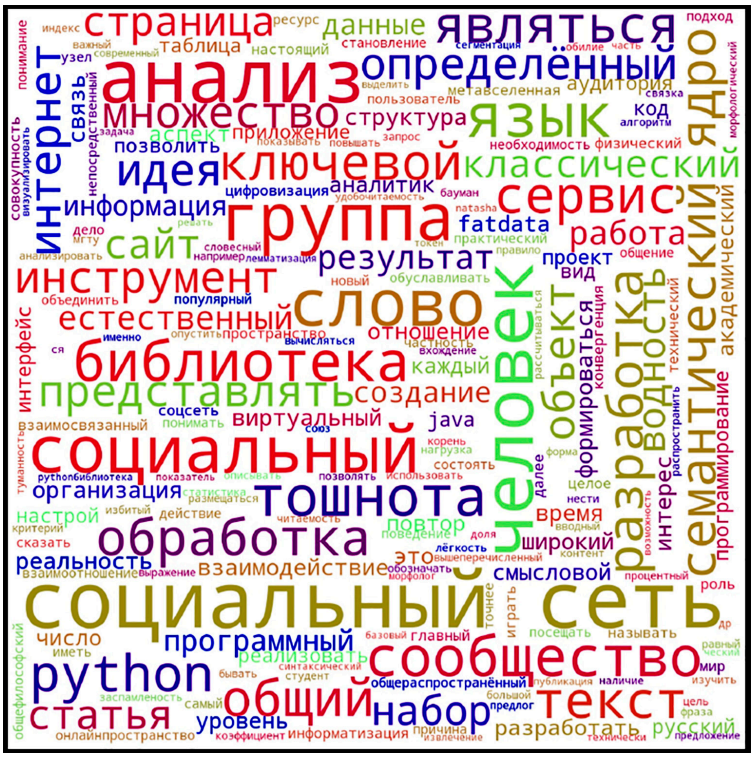


*Рис. 1. Облако слов*

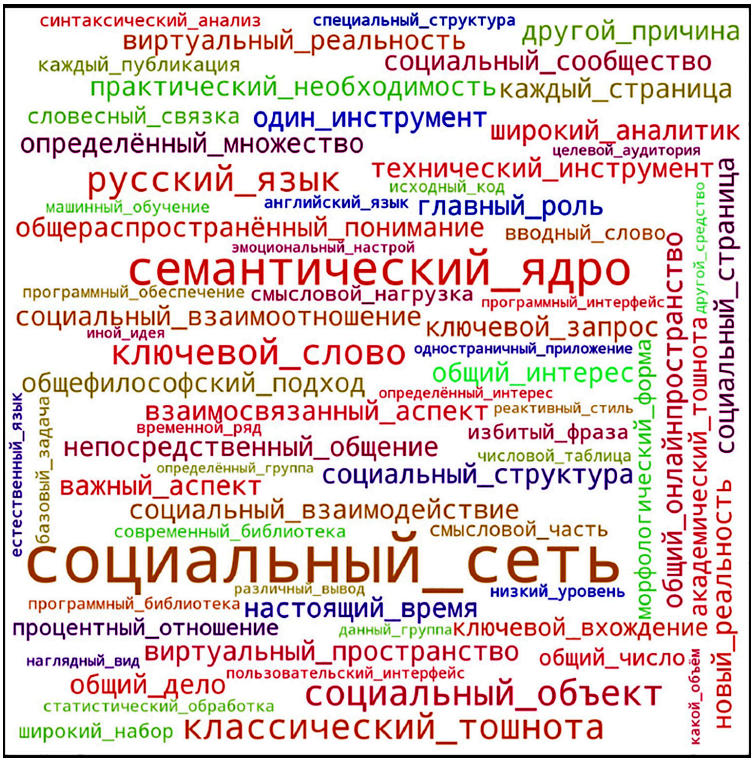


*Рис. 2. Облако биграмм*

С помощью метода make_agree_with_number выполняется согласование слова с числительным, например 20: w.make_agree_with_number(20).word. Результат выполнения метода: ‘аналитических’.

Для закрепления полученных знаний студенты выполняют практическое задание: с помощью морфологического анализатора текста Pymorphy2 выполнить анализ предложения объемом не менее 10 слов.

На занятии «Применение морфологического анализатора Pymorphy2 при решении прикладных задач» студенты составляют программу на языке python, в которой подключают библиотеку NLTK, загружают текст научной статьи из txt-файла, выполняют токенизацию текста.

Применив функцию частотного анализа текста, обучающийся печатает 10 токенов в порядке убывания частоты:

```
fdist = FreqDist(text)
print(fdist.most_common(10))
```

Результат выполнения кода не содержит никакой полезной информации:

[(‘и’, 29), (‘в’, 21), (‘для’, 14), (‘–’, 13), (‘из’, 9), (‘с’, 8), (‘на’, 8), (‘людей’, 7), (‘социальных’, 7), (‘социальные’, 6)].

Необходимо изучить методы применения стоп-слов и с помощью морфологического анализатора перевести токены в нормальную форму. В итоге программный код составляет последовательность ключевых слов, которые несут смысловую нагрузку, например: [(‘социальный’, 17), (‘сеть’, 10), (‘анализ’, 9), (‘человек’, 8), (‘группа’, 7), (‘слово’, 7), (‘язык’, 6), (‘сообщество’, 5), (‘тошнота’, 5), (‘библиотека’, 5)].

С помощью библиотеки WordCloud можно графически отобразить результаты частотного анализа текста в форме «облака слов» (рис. 1).

С помощью морфологического анализатора текста Pymorphy2 студенты осваивают методы извлечения биграмм (прилагательное + существительное) (рис. 2), триграмм (существительное + глагол + существительное) и т.п.

На занятии «Практическая работа на заданную тему» студент самостоятельно выполняет научно-исследовательскую работу по обработке и семантическому анализу текста, результатом которой становится «Отчет о НИР». Отчет должен содержать следующую информацию: источники исходных данных (электронные научные журналы, тезисы студенческих конференций), программы загрузки и предварительной обработки текстовых файлов, программы исследования текстов с использованием морфологического анализатора, визуализация результатов частотного анализа.

## Заключение

Разработана методика преподавания основ ОЕЯ, основанная на языке программирования Python и библиотеках Pymorphy2, NLTK и Wordcloud. Приобретенные знания дают студентам возможность приступить к изучению таких актуальных тем применения искусственного интеллекта, как извлечение фактов из неструктурированных текстов с помощью парсеров, основанных на грамматиках и правилах, оценка семантической близости документов, тематическое моделирование, построение графов и сетей социальных групп, поиск информационных кластеров и др.

Предложенный автором курс практических занятий прошел пилотную апробацию в Московском государственном техническом университете им. Н.Э. Баумана при подготовке бакалавров по направлениям 09.03.01 «Информатика и вычислительная техника» и 09.03.03 «Прикладная информатика» (профиль «Информационная аналитика»). На занятиях с помощью морфологического анализатора были исследованы статьи из девяти сборников студенческих работ «Аналитические технологии в социальной сфере: теория и практика», которые были представлены на ежегодных конференциях АНО «Научно-исследовательский центр проблем национальной безопасности». Составлены ключевые слова к ним, проведен семантический анализ изменения сферы научных интересов авторов статей за период с 2019 по 2023 г. Полученные результаты наглядно продемонстрированы в форме «облаков слов».

## Список литературы

1. Указ Президента РФ от 10 октября 2019 г. № 490 «О развитии искусственного интеллекта в Российской Федерации» (с изм. и доп.) [Электронный ресурс]. URL: https://base.garant.ru/72838946/ (дата обращения: 11.01.2025).

2. Оганесян С.А. Применение искусственного интеллекта в компьютерной лингвистике и обработке естественного языка // Вестник науки. 2024. № 7. С. 272–279. URL: https://www.вестник-науки.рф/article/16866 (дата обращения: 25.01.2025).

3. Полицына Е.В., Полицын С.А., Поречный А.С., Рыкунов А.Н. Анализ качества работы и расширение возможностей инструментов морфологического анализа текстов на русском языке // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2023. № 2. С. 171–180.
DOI: 10.17308/sait/1995-5499/2023/2/171-180.

4. Шульман В.Д., Максименко О.Е., Волхонцева П.Д. Анализ программных средств морфологического анализа // Международный журнал гуманитарных и естественных наук. 2022. № 3–2. С. 166–170.
DOI 10.24412/2500-1000-2022-3-2-166-170.

5. Приказ Министерства образования и науки РФ от 19 сентября 2017 г. № 929 «Об утверждении федерального государственного образовательного стандарта высшего образования – бакалавриат по направлению подготовки 09.03.01 Информатика и вычислительная техника» (с изм. и доп.) [Электронный ресурс]. URL: https://base.garant.ru/71784846/ (дата обращения: 11.01.2025).

6. Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages // Analysis of Images, Social Networks and Texts. 2015. P. 320–332. DOI:10.1007/978-3-319-26123-2_31.