УДК: 330:519.712.1
DOI:10.30987/2658-6436-2021-2-19-23

А.В. Иванова, Р.А. Филиппов, Л.Б. Филиппова,
А.С. Сазонова, А.А. Кузьменко, Ю.А. Леонов

ИССЛЕДОВАНИЕ МЕТОДОВ ОБРАБОТКИ ТЕКСТОВОЙ ИНФОРМАЦИИ И ОБЗОР ЭТАПОВ СОЗДАНИЯ МОДЕЛИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ПРИ СОЗДАНИИ ЧАТ-БОТОВ

Задача создания чат-бота на основе нейронной сети предполагает машинную обработку текста, которая в свою очередь подразумевает использование различных методов и способов анализа фраз и предложений. В статье рассмотрены наиболее популярные решения и модели анализа данных в текстовом формате: способы лемматизации, векторизации, а также методы машинного обучения. Особое внимание уделено методикам обработки текста, в последствии анализа которых был выявлен и протестирован наилучший метод.

Ключевые слова: *чат-бот, нейронная сеть, текст, лемматизация, векторизация, машинное обучение.*

A.V. Ivanova, R.A Filippov., L.B. Filippova, A.S. Sazonova, A.A. Kuzmenko, Yu.A. Leonov

RESEARCHING METHODS FOR PROCESSING TEXT INFORMATION AND REVIEWING THE STAGES OF AN ARTIFICIAL INTELLIGENCE MODEL CREATION AT PRODUCING CHATBOTS

The task of producing a chatbot based on a neural network supposes machine processing of the text, which in turn involves using various methods and techniques for analyzing phrases and sentences. The article considers the most popular solutions and models for data analysis in the text format: methods of lemmatization, vectorization, as well as machine learning methods. Particular attention is paid to the text processing techniques, after their analyzing the best method was identified and tested.

Keywords: *chatbot, neural network, text, lemmatization, vectorization, machine learning.*

Введение

В современном мире бизнес всегда должен оставаться конкурентноспособным и быть на голову впереди своих соперников. Здесь на помощь приходит технологическое развитие.

Помощники с искусственным интеллектом предоставляют возможность предпринимателям заметно экономить на времени и средствах как в общении с клиентами, так и в решении внутренних корпоративных задач. Современные чат-боты выходят на новый уровень и становятся все более схожи с человеческим разумом.

Возможность генерации уникальных ответов в чат-ботах осуществляется при помощи моделей MachineLearning и внедрении их в код программы. На данный момент нет четко отработанной методики создания чат-ботов с искусственным интеллектом, ведь для разных задач и возможностей применяются разные методы.

Исследование

Этапы создания модели искусственного интеллекта (рис. 1) нужны для приведения текстовых данных к форме, в которой анализирование слов становится возможным, ведь текст в своем исходном виде нельзя использовать в математической модели.

Для осуществления этапов создания модели ИИ, существует множество инструментов,

подходящих для классификации текста. Далее рассмотрены наиболее эффективные механизмы, подходящие для данной задачи.

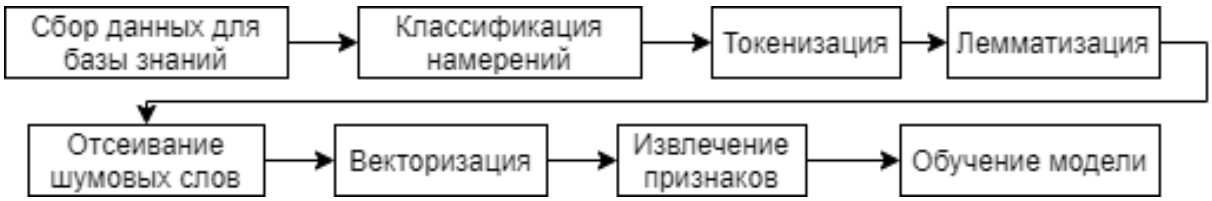

Рис. 1. Этапы создания чат-бота на основе искусственного интеллекта

1. Токенизация. Этот этап осуществляется посредством встроенных библиотек в актуальных языках программирования. Например, в языке Python имеются стандартные библиотеки: re (модуль для работы с функциональными выражениями) и string (модуль для работы с массивами и функциями, необходимыми для анализа строк)
2. Лемматизация. В базовых библиотеках языков программирования нет возможности осуществления лемматизации, но есть сторонние библиотеки в свободном распространении. Для языка Python наиболее удобными и популярными являются: NLTK, SpaCy и PyMorphy.
3. Удаление стоп-слов. Данный этап подразумевает осуществление алгоритма удаления ненужных слов (союзы, предлоги, междометия и т.д.) из имеющегося массива, идентифицированных как шум. Использование библиотек не требуется.
4. Векторизация и машинное обучение. Самым ключевым моментом в создании чат-бота на основе ИИ, является подбор и разработка алгоритма векторизации и машинного обучения, ведь различные методы подходят для разных нужд. Так как машина «не понимает» текст, то нужно предоставить данные для обработки в наиболее доступном формате – числах, которые необходимо сформировать из текстов, хранящихся в базе знаний. Для этой задачи и используется векторизация. Наиболее популярными методами являются: CountVectorizer и TF-IDF.

CountVectorizer (рис. 2) является самым простым и действенным способом представления текстовых документов. В данном методе токенизируются входные фразы и строится словарь известных слов. Если рассмотреть подробнее суть метода, то получается, что в раннее сформированной обучающей базе знаний все тексты разбиваются на уникальные проиндексированные слова и формируется словарь, из которого в дальнейшем можно составить любую фразу. Входная фраза переводится в вектор следующим образом: присваивается 1, если слово встречается в словаре и 0, если нет (таблица 1). По сути, полученные вектора обозначают количество слов, встретившихся в предложении. Таким способом можно выявить принадлежность входного предложения к тому или иному интенту (классу) из базы знаний.

Таблица 1. Векторное представление слов

	какой	**кино**	**посмотреть**	**кинотеатр**	**…**
какое кино посмотреть	1	1	1	0	…
где посмотреть кино	0	1	1	0	…
в каком кинотеатре посмотреть кино	1	1	1	1	…
…	…	…	…	…	…
как дела	0	0	0	0	…

```
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer()
X = vectorizer.fit_transform(X_texts)
```

Рис. 2. Использование метода векторизации CountVectorizer

Альтернативным и более сложным способом векторизации является метод TF-IDF (рис. 3). Суть та же самая, что и в предыдущем методе, но здесь добавляется вычисление частот слов. То-есть данный метод помогает отразить важность слова в документе. В основе лежат показатели:

TF – показывает насколько часто то или иное слово фигурирует в словаре (может быть посчитан методом CountVectorizer).

IDF – показатель обратной частоты документа, который снижает вес слова, часто встречающегося во всем словаре (предлоги, союзы, общие термины и понятия).

```
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(X_texts)
```

Рис. 3. Использование метода векторизации TF-IDF

Общая формула показателя IDF выглядит следующим образом:

$$IDF(t, D) = log \frac{|D|+1}{DF(t,D)+1} \quad (1)$$

Где D — количество документов в корпусе, DF(t, D) — количество документов, в которых встречается слово. Так, если слово встречается во всех документах, то IDF = 0. В итоге,

$$TFIDF = IDF * TF \quad (2)$$

Предположительно, этот способ должен работать лучше благодаря расчету веса слов, позволяющему получать релевантные результаты при минимально затраченном машинном ресурсе.

Тестирование

Чтобы выбрать наиболее подходящую модель обработки текстовой информации, необходимо произвести тестирование каждой из них и оценить результат.

Базу знаний следует разделить на две части: обучающую и тестируемую. Просто поделить пополам нельзя, иначе потеряется суть интентов (классов). Для этой задачи подойдет метод train_test_split (рис. 4), который правильно разделит массивы. Вручную зададим размер частей: 1/3 – тестовая выборка и 2/3 – обучающая выборка. При каждом запуске тестовой программы, значения будут получаться разные. Чтобы минимизировать неточности, создадим цикл запуска программы 10 раз, а затем посчитаем усредненное значение.

Также в тестировании используется классификатор скор – вероятность предсказания. Он берет очередной элемент из X и определять с помощью метода predict его класс, а затем сравнивает с соответствующим значением в Y и смотрит совпадение. Затем он делит сумму совпадений на общее количество и дает среднее значение, которое и является показателем качества обучившейся модели.

```
from sklearn.model_selection import train_test_split
```

Рис. 4. Метод тестирования train_test_split

Для обработки векторизации и построения зависимостей между текстами, необходимо разработать алгоритм обучения нейронной сети. Для поставленной цели отлично подойдет задача кластеризации. Из наиболее актуальных способов был выбран LogisticRegrassion.

Логистическая регрессия представляет собой метод для анализа данных, в которых присутствуют независимые переменные, определяющие результат (рис. 5). Применяется для прогнозирования двоичного результата: 1 и 0, да и нет, и т.д. с учетом набора независимых переменных.

```
from sklearn.linear_model import LogisticRegression
clf = LogisticRegression().fit(X, y)
```

Рис. 5. Использование метода машинного обучения LogisticRegrassion

Результат использования логистической регрессии и метода векторизации CountVectorizer представлен на рис. 6. По такому же принципу протестируем оставшуюся комбинацию с методом TF-IDF. Результаты приведены в таблице 2.

```
scores = []

for i in range(10):
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33)
    clf = LogisticRegression().fit(X_train, y_train)
    score = clf.score(X_test, y_test)
    scores.append(score)

sum(scores) / len(scores)
```

```
0.24648241206030147
```

Рис. 6. Использование LogisticRegression и CountVectorizer

Таблица 2. Качество обучения нейронной сети при различных комбинациях методов машинного обучения и векторизации

	CountVectorizer	**TF-IDF**
LogisticRegression	24,64 %	29,51 %

Выводы

Исходя из проведенного эксперимента, можно сделать вывод, что при обучении нейронной сети, наилучшим методом векторизации текста является TF-IDF. Показатель качества обучения составил 29,51 %. Из этого следует, что с вероятностью 29,51 % искусственный интеллект сможет ответить пользователю правильно на поставленный вопрос.

Показатель качества можно улучшить различными способами: подобрать и протестировать другие методы машинного обучения (например, LinearSVC), расширить базу знаний и провести более тщательную ее обработку.

Список литературы:

1. Бородин А. И., Вейнберг Р. Р., Литвишко О. В. , Методы обработки текста при создании чат-ботов / Хуманитарни Балкански изследвания, № 3(5) 31.08.2019

2. Филиппов, Р.А. Интернет вещей: основные понятия/ учебно-методическое пособие / Р.А. Филиппов, Л.Б. Филиппова, А.С. Сазонова — Брянск: БГТУ, 2016. — 112 с. — ISBN 978-5-906967-62-6 — Текст : непосредственный.

3. Leonov, YU.A. Selection of rational schemes automation based on working synthesis instruments for

References:

1. Borodin A.I., Veinberg R.R., Litvishko O.V., Methods of Text Processing when Creating Chatbots / Humanities Balkan Research, no. 3 (5) 31.08.2019

2. Filippov, R.A. Internet of Things: Basic Concepts / teaching aid / R.A. Filippov, L.B. Filippova, A.S. Sazonova. – Bryansk: BSTU, 2016. – 112 p. – ISBN 978-5-906967-62-6 - Text: unmediated

3. Leonov, Yu.A. Selection of Rational Schemes Automation Based on Working Synthesis Instruments for

technological processes / YU.A. Leonov, E.A. Leonov, A.A. Kuzmenko, A.A. Martynenko , E.E. Averchenkova, R.A. Filippov — Yelm, WA, USA: Science Book Publishing House LLC, 2019 — 192 p. — ISBN: 978-5-9765-4023-1 — Text : unmediated.

4. Leonov, E.A. Intellectual subsystems for collecting information from the internet to create knowledge bases for self-learning systems / E.A. Leonov, Y.A. Leonov, Y.M. Kazakov, L.B. Filippova/ In: Abraham A., Kovalev S., Tarassov V., Snasel V., Vasileva M., Sukhanov A. (eds) — Text : electronic // Proceedings of the Second International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'17). IITI 2017. Advances in Intelligent Systems and Computing. — 2017— vol 679. — Springer, Cham, p. 95-103 — DOI:10.1007/978-3-319-68321-8_10

5. Тищенко, А.А. Анализ конструкторов, позволяющих создавать мобильные приложения с целью развития цифровых технологий в логистических системах / А.А. Тищенко, Ю.М. Казаков – Текст непосредственный // X всероссийская научно-практическая конференция "Цифровая логистика - интегрированный подход" – 2020. – Брянск, БГТУ, с 181-185 - ISBN 978-5-907271-61-6.

Technological Processes / Yu.A. Leonov, E.A. Leonov, A.A. Kuzmenko, A.A. Martynenko, E.E. Averchenkova, R.A. Filippov – Yelm, WA, USA: Science Book Publishing House LLC, 2019 – 192 p. – ISBN: 978-5-9765-4023-1 – Text : unmediated.

4. Leonov E.A., Intellectual Subsystems for Collecting Information from the Internet to Create Knowledge Bases for Self-learning Systems / E.A. Leonov, Y.A. Leonov, Y.M. Kazakov, L.B. Filippova/ In: Abraham A., Kovalev S., Tarassov V., Snasel V., Vasileva M., Sukhanov A. (eds) – Text : electronic // Proceedings of the Second International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'17). IITI 2017. Advances in Intelligent Systems and Computing. –2017 – vol. 679. – Springer, Cham, pp. 95-103 – DOI:10.1007/978-3-319-68321-8_10

5. Tishchenko, A.A. Analysis of Constructors that Allow Creating Mobile Applications for Developing Digital Technologies in Logistics Systems / A.A. Tishchenko, Yu.M. Kazakov – Text : unmediated // Xth All-Russian Scientific and Practical Conference "Digital Logistics – Integrated Approach" – 2020. –Bryansk, BSTU, pp. 181-185 – ISBN 978-5-907271-61-6.

Статья поступила в редколлегию 28.04.2021.
Рецензент:
д-р. техн. наук, доц., Брянский государственный технический университет
Аверченков А.В.
Статья принята к публикации 13.05.2021.

Сведения об авторах:

Иванова Анастасия Владимировна
магистр кафедры «Компьютерные технологии и системы» Брянского государственного технического университета

Кузьменко Александр Анатольевич
кандидат биологических наук доцент Брянского государственного технического университета
E-mail: alex-rf-32@yandex.ru

Филиппов Родион Алексеевич
кандидат технических наук, доцент Брянского государственного технического университета
E-mail: redfil@mail.ru

Филиппова Людмила Борисовна
кандидат технических наук, доцент Брянского государственного технического университета
E-mail: libv88@mail.ru

Сазонова Анна Сергеевна
кандидат технических наук, доцент Брянского государственного технического университета
E-mail: asazonova@list.ru

Леонов Юрий Алексеевич
кандидат технических наук, доцент Брянского государственного технического университета
E-mail: yorleon@yandex.ru

Information about authors:

Ivanova A.V.
Master of the Department of Computer Technologies and Systems of Bryansk State Technical University

Kuzmenko A.A.
Candidate of Biological Sciences, Associate Professor of Bryansk State Technical University
E-mail: alex-rf-32@yandex.ru

Filippov R.A.
Candidate of Technical Sciences, Associate Professor of Bryansk State Technical University

Filippova L.B.
Candidate of Technical Sciences, Associate Professor of Bryansk State Technical University
E-mail: libv88@mail.ru

Sazonova A.S.
Candidate of Technical Sciences, Associate Professor of Bryansk State Technical University
E-mail: asazonova@list.ru

Leonov Yu.A.
Candidate of Technical Sciences, Associate Professor of Bryansk State Technical University
E-mail: yorleon@yandex.ru