

СОЗДАНИЕ ДИАЛОГОВОГО АССИСТЕНТА С ИСПОЛЬЗОВАНИЕМ КЛАССИФИКАТОРОВ В ОБРАБОТКЕ ЕСТЕСТВЕННОГО ЯЗЫКА

CREATION OF A DIALOGUE ASSISTANT USING CLASSIFIERS IN NATURAL LANGUAGE PROCESSING

N. Verezubova
N. Sakovich
I. Verezubova
A. Chekulaev

Summary. The article proposes an optimal architecture for a chatbot based on classifier models, which includes key components (natural language processing, machine learning and integration with databases). The classification models used in chatbots are considered, which make it possible to accurately determine the intentions of users and adapt responses in accordance with their requests, as well as the possibility of their training based on user interaction.

Keywords: chatbot, virtual assistant, machine learning, classifiers, stochastic gradient descent, random forest classifier.

Вереzubова Наталья Афанасьевна

Кандидат экономических наук, доцент, Московская государственная академия ветеринарной медицины и биотехнологии имени К.И. Скрябина
nverez@mail.ru

Сакович Наталия Евгениевна

Доктор технических наук, доцент, Брянский государственный аграрный университет
nasa2610@mail.ru

Вереzubова Ирина Николаевна

Московский государственный педагогический университет

Чекулаев Артур Анатольевич

Московская государственная академия ветеринарной медицины и биотехнологии имени К.И. Скрябина

Аннотация. В статье предлагается оптимальная архитектура для чат-бота, основанного на классификаторных моделях, включающая в себя ключевые компоненты (обработка естественного языка, машинное обучение и интеграция с базами данных). Рассматриваются классификаторные модели, используемые в чат-ботах, которые позволяют точно определять намерения пользователей и адаптировать ответы в соответствии с их запросами, а также возможностью их обучения на основе пользовательского взаимодействия.

Ключевые слова: чат-бот, виртуальный ассистент, машинное обучение, классификаторы, стохастический градиентный спуск, классификатор случайного леса.

Введение

В современном мире, бурно развивающимся и полным инноваций, машинное обучение играет ключевую роль в усовершенствовании технологий, облегчающих жизнь человека.

Машинное обучение — это общее название для группы алгоритмов, имитирующих процессы человеческого мышления, оно включает в себя, как и нейронные сети, так и задачи регрессии и классификации [1].

Данный кластер моделей может быть использован для создания множества различных программ и систем, в том числе и виртуальных ассистентов [2, 3].

Виртуальных ассистентов можно классифицировать по признаку использования в них нейросетевых алгоритмов, разделяя принципы работы чат ботов на две группы — с использованием нейросетевых алгоритмов и без их применения, и по принципу диалоговой направленности программ, разделяя их на три категории — FAQ-чат-бот, целенаправленный чат-бот, chit chat bot [4, 5].

FAQ-чат-бот это одна из самых простых архитектур виртуальных помощников, основанная на искусственном интеллекте, включающая в себя NLP (Natural language processing) — обработку естественного языка. Этот тип ассистентов автоматизирует поддержку клиентов различных компаний, использующих данных ботов [6].

Целенаправленный чат-бот — это помощник, основанный на нейронной сети, натренированной решать конкретные задачи, связанные с взаимодействием с клиентами. Данный тип ассистентов помогает компаниям найти персонализированный подход к каждому клиенту [7].

Чат-боты вопросно-ответной системы (chit chat bot) — это система, извлекающая информацию для разговора из какого-либо источника, например, текстового документа с заданными фразами формата «Вопрос пользователя — 'это' ответ бота» [8].

Постановка задачи

Чат боты могут основываться не только на нейронных сетях, но и на классификации данных. Данные алго-

ритмы в теории могут быть менее ресурсозатратными и при должном обучении являться моделью, помогающей другим системам обрабатывать текст [9].

Чтобы исследовать возможности подобных моделей, было решено смоделировать чат-бот, реализующий в себе разные модели классификации, в качестве архитектур были выбраны классификатор опорных векторов (SVM) с оптимизацией в виде стохастического градиентного спуска (SGD) и классификатор случайного леса (RF). SVM это метод классификации, основанный на разделении выборок (векторов) на классы при помощи условной гиперплоскости [10].

На рисунке 1 показана схема работы классификатора опорных векторов.

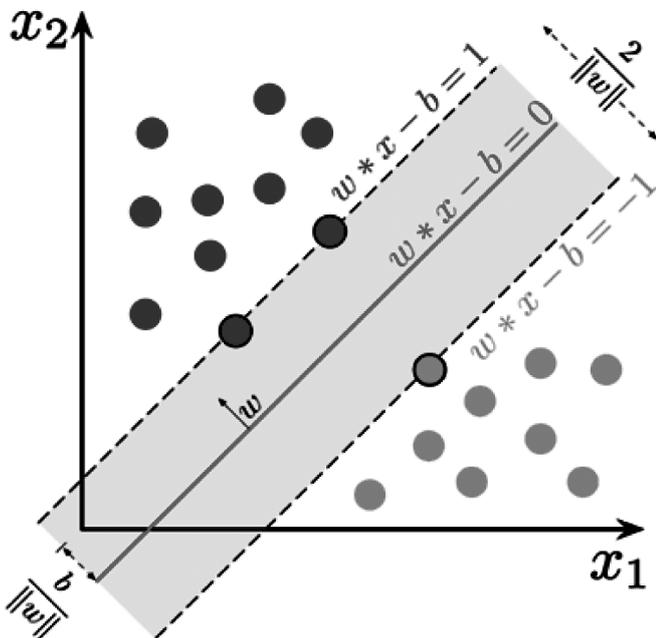


Рис. 1. Схема работы классификатора опорных векторов

В свою очередь классификатор случайного леса — это ансамблевый метод классификации, то есть система, состоящая из множества подсистем, действующих вместе, в данном случае это совмещение двух методов — Бэггинга и метода случайных пространств над решающими деревьями [11].

Методы исследования

Основным методом исследования выступит моделирование системы на базе машинного обучения. В качестве системы выступит чат-бот обученный на определённой выборке данных.

Ассистент обучается при помощи двух, ранее представленных систем, настроенных под данную выборку данных. На рисунке 2 представлен фрагмент кода, определяющий тонкую настройку модели на выборке

из 1709 фраз, представленных в виде четырёх словарей, с данными формата «Вопрос пользователя=Ответ бота» и «Вопрос пользователя=Ответ бота | Вариация ответа бота», данные для обучения были сгенерированы искусственно при помощи нейросетевых моделей для генерации текста.

```

model1 = SGDClassifier(max_iter=1000, loss="hinge",
    ..... random_state=72)
model1.fit(X_train, y_train)

# Модель случайного леса
model2 = RandomForestClassifier(random_state=73,
    ..... min_samples_split=4,
    ..... min_samples_leaf=2
    ..... )
model2.fit(X_train, y_train)
    
```

Рис. 2. Фрагмент кода

В качестве контекста диалога была применена реализация метода, подсчёта энтропии Шеннона для классификации намерений. Данный метод был разработан как альтернатива классическим моделям векторного представления [12].

Данные, поступившие из словаря кластеризируются, при помощи алгоритма к средних ++, из библиотеки scikit-learn, и поступают в контекстное окно. Бот устроен таким образом, что при вводе пользователем фразы, происходит подсчёт энтропии Шеннона не только ввода человека, но и вывода бота, и при слишком высокой, как и при слишком низкой энтропии на выводе, бот получает штраф к весам модели, затем сопоставляя штраф и энтропию ввода, корректирует веса, представленные общим «скором» (score) выходные фразы с учётом штрафов.

На рисунке 3 показана схема работы чат бота.

Для усиления понимания ассистентом контекста, был применён расчёт косинусного расстояния между векторизованными участками выбранных ботом из словарей паттернов для ответа, если же ответ не найден, то бот просит пользователя обучить его новой фразе.

Результаты и их обсуждение

В ходе ведения диалога ассистентом было выяснено, что данная архитектура достаточно гибкая и в теории может применяться в поддержке таких моделей как FAQ и chat чат-боты, контролируя их контекст при помощи окна с контролем диалога при помощи штрафа за энтропию.

Однако у данной модели есть и недостатки, она крайне чувствительна к переобучению, из-за чего при помощи sklearn.model_selection.train_test_split, приходится разделять кластеризованные данные на несколько выборок, также обязательным стало случайное перемешивание массивов при помощи shuffle=True, на рисунке

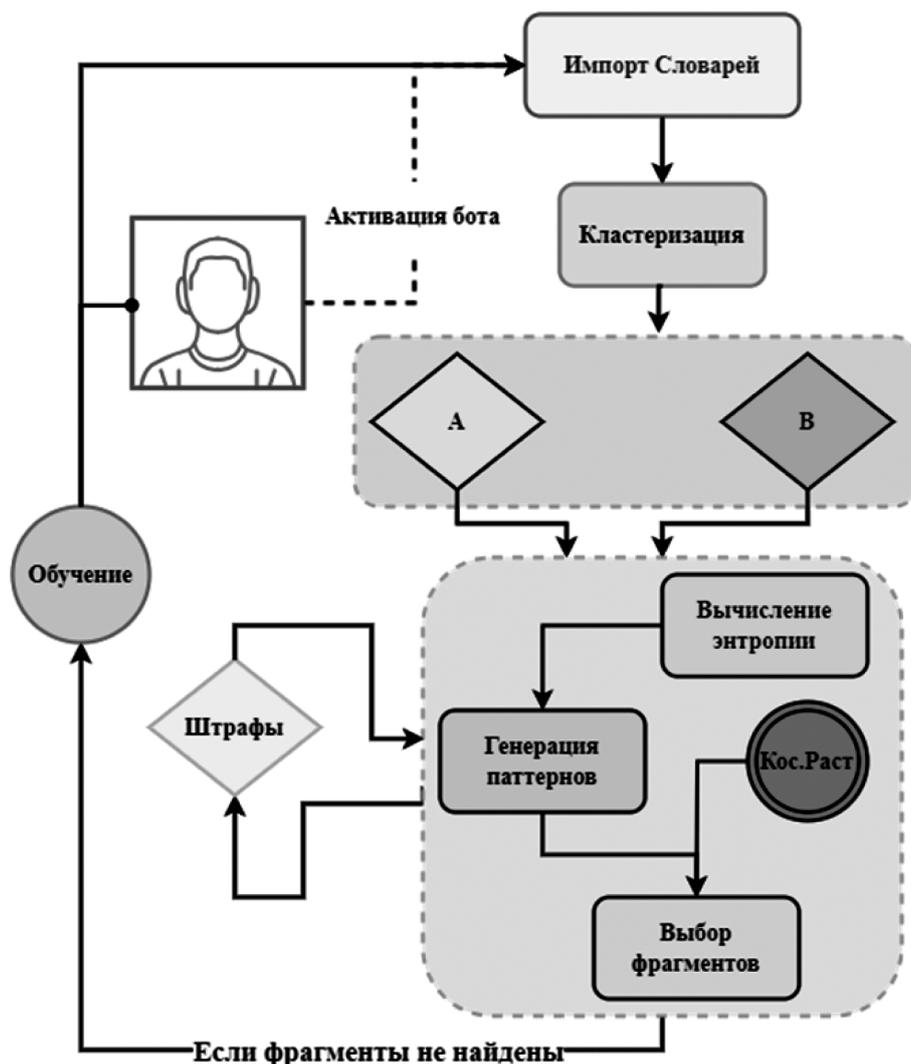


Рис. 3. Схема работы чат бота

ке 4, представлен фрагмент кода, реализующий данную функцию.

```
X_train, X_test, y_train, y_test = train_test_split(
    vectorizer.transform(
        cluster_questions,
        cluster_responses,
        test_size=0.2,
        random_state=72,
        train_size=0.8,
        shuffle=True))
```

Рис. 4. Фрагмент кода

Можно сделать вывод, что чат-боты, основанные на не нейросетевых алгоритмах, в данном случае на алгоритмах классификации, могут выполнять функции виртуального помощника, однако, их архитектура остаётся нестабильной, а обучение данной модели будет строится в основном на подборе весов классификатора в связи с его чувствительностью к переобучению, а также на мануальном вводе фраз для чат бота в словари, что при определённых обстоятельствах может оказаться крайне неудобным и не эргономичным. Однако элемен-

ты данных моделей могут выполнять функцию поддержки, улучшая понимание контекста у не продвинутых нейросетевых алгоритмов. Подобные алгоритмы при относительной простоте своей реализации, остаются лишь экспериментальными и нуждаются в осмыслении и доработке.

Заключение

В ходе работы нами была создана и обучена модель виртуального ассистента. Основной целью исследования было изучение гибкости архитектуры ассистента и его способности к естественному общению в диалоге. Для оценки коммуникативных навыков был проведен ряд тестов, включающих в себя различные сценарии взаимодействия с ассистентом.

Результаты исследования показали, что разработанная модель обладает определенной степенью гибкости, что позволяет ей следовать чётко поставленной модели диалога. Она способна узнавать ключевые слова, анали-

зировать структуру предложения и давать более-менее адекватные ответы. Однако, мы также заметили ряд серьезных ограничений. Во-первых, модель продемонстрировала высокую чувствительность к переобучению. Во-вторых, модель отличалась нестабильностью. В-третьих, она не способна адаптироваться к мульти контекстным задачам.

С учетом выявленных недостатков следует сделать вывод о том, что разработанная модель виртуального ас-

систента требует значительной доработки. Необходимо разработать новые методы обучения, которые помогли бы предотвратить переобучение и увеличить стабильность модели. Также нужно рассмотреть использование других архитектур и методов, которые могли бы предоставить более устойчивые и предсказуемые результаты. В целом, наша работа показывает, что создание виртуальных ассистентов основанных на классификаторных моделях остаётся не полностью изученной, но перспективной темой для дальнейших исследований.

ЛИТЕРАТУРА

1. Митрофанова А.С. Машинное обучение / А.С. Митрофанова, Г.В. Комлев // Тенденции развития науки и образования. — 2018. — № 44–7. — С. 39–40.
2. A comprehensive guide to chatbot machine learning. [Электронный ресурс]. — Режим доступа: <https://cyfuture.com/blog/a-comprehensive-guide-to-chatbot-machine-learning/>. (16.09.2024).
3. Chatbot with Machine Learning and Python / [Электронный ресурс]. — Режим доступа: <https://thecleverprogrammer.com/2020/11/01/chatbot-with-machine-learning-and-python/>. (16.09.2024).
4. Что внутри чат-бота? [Электронный ресурс]. — Режим доступа: <https://habr.com/ru/companies/mipt/articles/469981/>. (16.09.2024).
5. Ураев, Д.А. Классификация и методы создания чат-бот приложений / Д.А. Ураев // International scientific review of the problems and prospects of modern science and education: COLLECTION OF SCIENTIFIC ARTICLES. LXIV INTERNATIONAL CORRESPONDENCE SCIENTIFIC AND PRACTICAL CONFERENCE, Boston, 20–21 ноября 2019 года / EDITOR: EMMA MORGAN. — Boston: PROBLEMS OF SCIENCE, 2019. — С. 30–33.
6. FAQ chatbots explained. [Электронный ресурс]. — Режим доступа: <https://www.sprinklr.com/cxm/faq-chatbots/>. (16.09.2024).
7. Goal-Oriented Bots: How to Elevate Customer Experience with Chatbots. [Электронный ресурс]. — Режим доступа: <https://www.tensorway.com/post/goal-oriented-bots>. (16.09.2024).
8. 5 методов обработки естественного языка, которые стремительно меняют мир вокруг нас / [Электронный ресурс]. — Режим доступа: <https://github.com/hightemp/docNN/blob/master/articles/5%20методов%20обработки%20естественного%20языка%20которые%20стремительно%20меняют%20мир%20вокруг%20нас.md>. (16.09.2024).
9. Исследование методов обработки текстовой информации и обзор этапов создания модели искусственного интеллекта при создании чат-ботов / А.В. Иванова, А.А. Кузьменко, Р.А. Филиппов [и др.] // Автоматизация и моделирование в проектировании и управлении. — 2021. — № 2(12). — С. 19–23.
10. Aung Z. The Support Vector Method for Solving the Classification Problem with Domain Constraints / Z. Aung, I.S. Mikhaylov, Y.T. Aung // Proceedings of the International Conference on Intelligent Computing and Control Systems, ICICCS 2020, Madurai, 13–15 мая 2020 года. — Madurai, 2020. — P. 32–36.
11. Ансамбли в машинном обучении. [Электронный ресурс]. — Режим доступа: <https://education.yandex.ru/handbook/ml/article/ansambli-v-mashinnom-obuchenii>. (21.07.2024).
12. Aleksandr Perevalov, Daniil Kurushin, Rustam Faizrakhmanov, Farida Khabibrakhmanova / Embeddings Based on Shannon Entropy Solving intent classification task in goal Doriented dialogue system / Aleksandr Perevalov, Daniil Kurushin, Rustam Faizrakhmanov, Farida Khabibrakhmanova // Proc. of the 7th International Conference on Applied Innovations in IT. — Perm: 2019. — С. 73–78.

© Везубова Наталья Афанасьевна (nvezub@mail.ru); Сакович Наталия Евгениевна (nasa261@mail.ru);

Везубова Ирина Николаевна; Чекулаев Артур Анатольевич

Журнал «Современная наука: актуальные проблемы теории и практики»