DOI 10.37882/2223–2966.2022.05.06

МЕТОДЫ АВТОНОМНОГО СБОРА И ОЦЕНИВАНИЯ КАЧЕСТВА ЛОКАЛЬНЫХ КОНТЕКСТОВ

METHODS OF AUTONOMOUS COLLECTION AND EVALUATION OF THE QUALITY OF LOCAL CONTEXTS

N. Vorobyev

Summary. The purpose of this study is to evaluate the methods of autonomous collection and use of local contexts to select an answer by an intelligent system in the process of dialog interaction. Methods of evaluating the quality of the system are given.

Keywords: DBMS, artificial intelligence, local context, dialog interaction, thesaurus, chatbot.

Воробьев Никита Григорьевич
Аспирант, Московский Политехнический университет, Москва
nickikta@yandex.ru

Аннотация. Целью данного исследования является оценка методов автономного сбора и использования локальных контекстов для выбора варианта ответа интеллектуальной системой в процессе диалогового взаимодействия. Приводятся методы оценивания качества работы системы.

Ключевые слова: СУБД, искусственный интеллект, локальный контекст, диалоговое взаимодействие, тезаурус, чат-бот.

В ходе научной деятельности передо мной встала задача разработки программного модуля, который встроен в код чат-бота, чтобы собирать локальный контекст во время диалога с пользователем и затем отвечать на запросы на его основе. Общаясь с пользователями определенного ресурса, чат-бот должен собирать контекст, присущий этой сфере деятельности, и использовать его в будущем. В открытом доступе находится множество предлагаемых способов решения отдельных задач, входящих в общий комплекс работы, но единой сравнительной характеристики не существует. Выяснению особенностей различных методов обработки данных и оцениванию качества и посвящена данная работа.

Схему функционирования разрабатываемого приложения можно представить в виде двух фаз. На первом этапе чат-бот выступает в роли стандартного ассистента: он подключается к ресурсу, разработчик выбирает запросы и ответы на них, вводит в базу данных, подключенную к боту, и активирует его. При этом приложение стандартно отвечает при общении с пользователями ресурса, но сохраняет слова и фразы, которые сопровождают заданный вопрос и последующий ответ. Эти термины формируют локальный контекст веб-ресурса.

На втором этапе, когда контекст сформирован, система упрощает полученные данные до ключевых структур. Используя собственную практику, бот обнаруживает целый ряд запросов, сводимых к ключевым словам, и выбирает на их основе более точный ответ. Затем он изменяет выбранные слова в ответе, чтобы лучше соответствовать ожиданиям пользователя.

Для самой системы предполагается создать три модуля, реализующие три метода анализа ввода пользователя: модуль статистического анализа, модуль тезаурусного поиска и модуль интеллектуального поиска. Для обеспечения взаимодействия между ними так же предназначен отдельный объект системы. Остановимся подробнее на каждом из реализованных методов анализа данных.

Методы распознавания запросов

Статистический анализ.

Статистический анализ является базовым методом, необходимым для сбора локального контекста [3]. Строка ввода пользователя, приходящая на вход, очищается от незначащих слов, и сравнивается со всеми сохраненными в базе данных вопросами. Если процент соответствия пересекает пороговое значение идентичности, то вопрос считается распознанным верно, и модуль возвращает идентификатор строки базы данных, содержащей ответ на вопрос пользователя.

Если процент соответствия пересекает только пороговое значение потенциальной идентичности, то пользователю задаются дополнительные вопросы, и на их основе осуществляется выбор подходящего ответа. Затем все слова, присутствовавшие во вводе пользова-

теля, и отличающиеся от соответствующих слов верно распознанного вопроса записываются в новую таблицу базы данных как контекстные синонимы. Данные в этой таблице хранятся в формате «слово» — «синоним» — «частота». При первой встрече определенной пары слов поле «частота» задается равной единице, и увеличивается на один с каждой последующей встречей. Затем данные из этой таблицы используются методом статистического анализа для повышения эффективности.

При обработке строки ввода вместе с чистым пользовательским вводом сохраняется массив альтернативных вводов, в котором хранятся строки ввода пользователя, в котором слова, присутствующие в таблице синонимов, заменены на свои аналоги. Таким образом вместо чистого ввода пользователя анализируется уже все семантическое поле, и вероятность распознавания значительно возрастает. Таблица, содержащая синонимы, и является локальным контекстом текущей сферы взаимодействия с системой.

Алгоритм работы:
1. Сравнивает заданный вопрос с каноничными;
2. При прохождении порога идентичности задает уточняющий вопрос;
3. Сохраняет альтернативные слова и формирует локальный контекст;
4. Использует локальный контекст для выбора ответа в последствии.

В таком случае система получится легкой в администрировании, но корректная настройка значений параметров переходов будет крайне затруднительной.

Тезаурусный поиск.

Второй метод — тезаурусный поиск [1]. Для его использования помимо обозначенных бах данных так же необходим тезаурус. Это специфичный файл, содержащий данные в формате «слово» — «лемма» — «частота», и другие поля. Для своей работы я использовал тезаурус, размеченный языком XML.

Сам модуль разбивает ввод пользователя, очищенный от незначащих слов, на двусловия, и обрабатывает уже их. Для каждого двусловия находится лемма, если ее нет для двусловия, то лемма находится для каждого отдельного слова. Массив лемм очищается от повторений, и сохраняется как обработанный ввод пользователя. Затем по такой же схеме обрабатываются все сохраненные в таблице вопросов и ответов вопросы, и сравниваются с результатом работы предыдущего шага. При пересечении порогового значения соответствия, вопрос считается распознанным и модуль возвращает ответ.

Алгорим работы:
1. Нормализует запрос;
2. Разделяет запрос на двусловия;
3. Извлекает из двусловий леммы;
4. Сравнивает леммы с леммами, соответствующими известным запросам.

При этом система будет сразу готова к работе и покажет достаточно стабильные результаты распознавания, но процесс функционирования будет значительно замедлен из-за множественных обращений к базам данных.

Интеллектуальный поиск

Он реализуется с помощью нейронной сети [7]. Данная сеть обучается на примерах подходящих и не подходящих вопросов. Ввод пользователя так же очищается от незначащих слов и разбивается на двусловия. Затем двусловия кодируются с помощью фэш-функции, и используются как параметры для ввода в нейросеть. На выход она возвращает числовые значения, сумма которых сравнивается с идентификаторами сохраненных вопросов. При успешном распознавании возвращается ответ.

Нейронная сеть обучается на подходящих и не подходящих вариантах вопросов, алгоритм работы:

1. Анализирует ввод и выдает ответ.

Такая система будет работать наиболее быстро из-за малого числа обращений к базе данных, но само обучение нейронной сети может занять продолжительное время, кроме того, для этого потребуется большой объем размеченных данных в обучающей выборке.

Все методы анализа можно объединить, чтобы ликвидировать недостатки. На старте работы запросы анализируются тезаурусным поиском, так как он дает самый стабильный результат. Параллельно статистический модуль собирает локальный контекст, который впоследствии заменит собой тезаурус. Распознанные и ошибочные запросы используются для обучения нейронной сети. Результат: В начале работы система будет отвечать достаточно медленно, но со временем заменит глобальный тезаурус локальным контекстом, а затем обучит нейронную сеть на собранных данных и скорость работы многократно возрастет.

Методы обработки слов

Bag of words

Слово кодируется частотой, с которой оно встречается в обучающей выборке. Это упрощенное представ-

$$H\left(\frac{1}{n}\right) = -\frac{1}{n}\ln\left(\frac{1}{n}\right) - \left(1-\frac{1}{n}\right)\ln\left(1-\frac{1}{n}\right);$$
$$M(n) = nH\left(\frac{1}{n}\right).$$

Рис. 1. Формула величины энтропии.

$$A\left(\frac{1}{n}\right) = -\frac{1}{n}\left(\ln\left(\frac{1}{n}\right)\right)^2 - \left(1-\frac{1}{n}\right)\left(\ln\left(1-\frac{1}{n}\right)\right)^2;$$
$$\sigma(n) = \sqrt{n\left[A\left(\frac{1}{n}\right) - \left(H\left(\frac{1}{n}\right)\right)^2\right]};$$

Рис. 2. Среднеквадратическое отклонение энтропии и величина коэффициента вариации тезауруса.

ление текста, которое используется в обработке естественных языков и информационном поиске. В этой модели текст (одно предложение или весь документ) представляется в виде мешка (мультимножества) его слов без какого-либо учета грамматики и порядка слов, но с сохранением информации об их количестве. Мешок слов обычно используется в методах классификации документов, где частотность вхождения слова используется как признак для обучения классификатора [4].

Проблема этого метода заключается в том, что для корректной работы необходимо, чтобы пользователь вводил слова без ошибок, так как слово, введенное с ошибкой, будет считаться отдельным.

Word2vec

Слово кодируется в вектор в трехмерном пространстве, расстояние по которым откладывается по нескольким параметрам.

Работа программы осуществляется следующим образом: word2vec принимает большой текстовый корпус в качестве входных данных и сопоставляет каждому слову вектор, выдавая координаты слов на выходе. Сначала он генерирует словарь корпуса, а затем вычисляет векторное представление слов, «обучаясь» на входных текстах. Векторное представление основывается на контекстной близости: слова, встречающиеся в тексте рядом с одинаковыми словами (а следовательно, имеющие схожий смысл), будут иметь близкие (по косинусному расстоянию) векторы. Полученные векторные представления слов могут быть использованы для обработки естественного языка и машинного обучения [5].

Для кодирования в W2V, ориентируясь на рассмотренные методы анализа, возможно использовать следующие показатели:

1 показатель — номер слова в тезаурусе, распределенном в алфавитном порядке;

2 показатель — количество синонимов;

3 показатель — число вхождений слова в тезаурус.

Оценка эффективности тезауруса

Основной критерий эффективности тезауруса — полная энтропия множества его элементов.

Критерий мощности тезауруса определяется его информационным потенциалом. Показателем может служить полная энтропия множества элементов. Пусть множество элементов тезауруса состоит из n элементов. Каждый элемент отождествляется только с одним существительным и не имеет ни одного прилагательного. Пусть вероятность выбора любого элемента из совокупности одинакова и равна 1/p = n. В этом случае средняя величина энтропии на один элемент и для всех элементов представлена на рисунке 1.

Величина энтропии будет зависеть не только от количества элементов, но и от основания логарифма. Для определённости будет использоваться натуральный алгоритм, единицей измерения при этом будет нит. Для аппроксимации величины мощности каким-либо распределением вероятности определим также второй начальный момент, среднеквадратическое отклонение энтропии и величину коэффициента вариации тезауруса (рисунок 2):

На рисунке 3 приведены графики функций (1)–(3). По виду кривых следует, что пара-

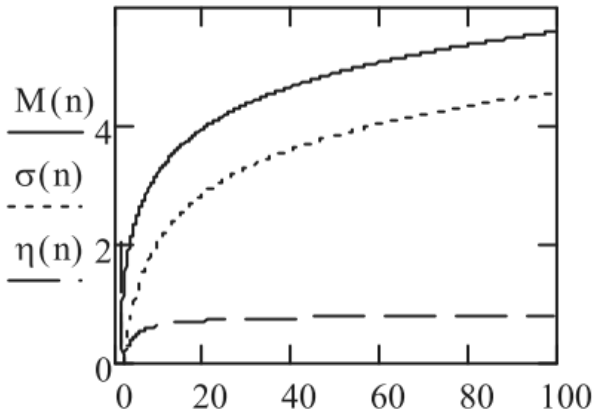

Рис. 3. Сравнение графиков функций.

$$p(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}}\exp\times$$

$$\times\left(-\frac{1}{2\sqrt{1-r^2}}\left(\frac{(x-a)^2}{\sigma_x^2}-2r\frac{(x-a)(y-b)}{\sigma_x\sigma_y}+\frac{(y-b)^2}{\sigma_y^2}\right)\right),$$

которая для суммирования приводится к формуле

$$p(x)=\frac{1}{\sqrt{2\pi\sigma_x^2+2r\sigma_x\sigma_y+\sigma_y^2}}e^{-\frac{(x-a-b)^2}{2(\sigma_x^2+2r\sigma_x\sigma_y+\sigma_y^2)}}.$$

Рис. 4. Формула двумерного распределения.

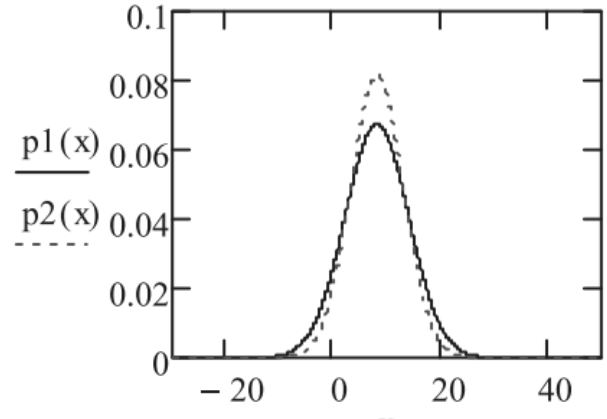

Рис. 5. Графики эффективности тезаурусов.

метры распределения случайной величины мощности увеличиваются с ростом количества элементов тезауруса n. Предельная величина коэффициента вариации монотонно стремится к η(∞) = 1. Для аппроксимации распределения случайной величины мощности тезауруса проще всего воспользоваться при η < 1 нормальным, а при η ≈ 1 — экспоненциальным законами распределения. Рассмотрим пример.

Пусть имеются два тезауруса, один с n = 10, второй — с n = 50. В соответствии с рисунком 1 параметры распределений будут равны: a = M(10) = 3,251; σ = σ(10) = 2,084;

b M= = (50) 4,902; σ=σ = (50) 3,853. Оба коэффициента вариации меньше единицы. Требуется объединить оба тезауруса в один для двух значений с коэффициентом корреляции r = 1 и r = 0,3. Для выполнения операции суммирования воспользуемся формулой для двумерного распределения (рисунок 4).

В данном случае для двух разных значений параметров и коэффициента корреляции можно представить их как p1(x) и p2(x). На рисунке 5 представлены графики функций плотности вероятностей и функций распределения суммарной мощности тезаурусов при указанных значениях параметров.

Рассмотренный в данном разделе тезаурус называется простым синтаксическим тезаурусом. Следует отметить, что такой тезаурус может быть расширен по горизонтали. Это можно объяснить тем, что его основные члены могут быть дополнены производными членами. Но и такой тезаурус остаётся также синтаксическим, хотя и становится более сложным, потому что дополнительные члены тезауруса по-прежнему являются, по Карнапу, не «прилагательными», а только «существительными» [2].

Таким образом, делая периодические срезы, и сравнивая эффективность тезаурусов подобным методом, можно балансировать динамические параметры методов распознавания, приведенных в начале исследования для улучшения качества работы.

Результаты

Исследованные в данной работе методы могут быть эффективно использованы для разработки саморегулируемой системы автономного обучения речевому взаимодействию с пользователями. Данный модуль будет оценивать эффективность своей работы и оптимизировать процесс взаимодействия с пользователем. Результаты данной работы можно будет применять в лингвистических исследованиях и корпоративной деятельности.

ЛИТЕРАТУРА

1. Бурый А.А., Зацерковный А.В., Поздняк П.Л. Система интеллектуального поиска на основе семантических сетей. URL: https://cyberleninka.ru/article/n/sistema-intellektualnogo-poiska-na-osnove-semanticheskih-setey/viewer (дата обращения: 11.01.2022).
2. Левченко С.В. Разработка метода кластеризации слов по смысловым характеристикам с использованием алгоритмов Word2Vec. URL: https://cyberleninka.ru/article/n/razrabotka-metoda-klasterizatsii-slov-po-smyslovym-harakteristikam-s-ispolzovaniem-algoritmov-word2vec/viewer (дата обращения: 11.01.2022).
3. Матвеева Н.Ю., Золотарюк А.В. Технологии создания и применения чат-ботов. // Научные записки молодых исследователей № 1/2018. URL: https://cyberleninka.ru/article/n/tehnologii-sozdaniya-i-primeneniya-chat-botov/viewer (дата обращения: 11.01.2022).
4. Мухин М.Ю. Психолингвистические аспекты лексико-статистического анализа текста. URL: https://elar.urfu.ru/bitstream/10995/95516/1/978-5-7996-3178-9_2021_007.pdf (дата обращения: 11.01.2022).
5. Проскурин А.А., Авсеева О.В. Объектно-ориентированная реализация обработки текста на основе алгоритма continuous bag of words. // Материалы V научно-практической конференции «ОЪЕКТНЫЕ СИСТЕМЫ — 2011» (Зимняя сессия). URL: https://cyberleninka.ru/article/n/obektno-orientirovannaya-realizatsiya-obrabotki-teksta-na-osnove-algoritma-continuous-bag-of-words/viewer (дата обращения: 11.01.2022).
6. Парамонов И.Ю., Смагин В.А. Мера информационной мощности тезауруса и её применение. // Интеллектуальные технологии на транспорте. 2016. № 4. URL: https://cyberleninka.ru/article/n/mera-informatsionnoy-moschnosti-tezaurusa-i-eyo-primenenie/viewer (дата обращения: 11.01.2022).
7. Журавлев А.П. Актуальность применения тезаурусного подхода для моделирования лексико-семантической структуры терминополя. // Социальные науки. 2021. URL: https://cyberleninka.ru/article/n/aktualnost-primeneniya-tezaurusnogo-podhoda-dlya-modelirovaniya-leksiko-semanticheskoy-strukturyterminopolya/viewer (дата обращения: 11.01.2022).

© Воробьев Никита Григорьевич (nickikta@yandex.ru).

Журнал «Современная наука: актуальные проблемы теории и практики»