УДК 004.03

Е.В. ИГИТЯН, К.А. ПОЛЬЩИКОВ

E. V. IGITYAN, K. A. POLSHCHYKOV

**ОБЗОР СРЕДСТ ОЦЕНИВАНИЯ ХАРАКТЕРИСТИК ДИАЛОГОВЫХ СИСТЕМ**

**OVERVIEW OF TOOLS FOR EVALUATING THE CHARACTERISTICS OF DIALOG SYSTEMS**

*В данной статье авторы освещают проблему отсутствия адекватных средств оценки качества диалоговых систем, которые бы удовлетворяли общепризнанные методы, модели, метрики, т.е. количественных показателей, на основе которых можно было бы сравнить эффективность имеющихся систем, обосновать выбор подходящих для тех или иных целей виртуальных ассистентов, а также объективно выявить их недостатки с целью дальнейшего совершенствования.*

*Ключевые слова: диалоговые системы, эффективность чат-бот-приложений.*

In this article, the authors highlight the problem of the lack of adequate means of assessing the quality of dialog systems that would satisfy generally accepted methods, models, metrics, i.e. quantitative indicators on the basis of which it would be possible to compare the effectiveness of existing systems, justify the choice of suitable virtual assistants for certain purposes, as well as objectively identify their shortcomings in order to further improvements.

Keywords: dialog systems, efficiency of chatbot applications.

В настоящее время диалоговые системы находят широкое применение в различных сферах деятельности, давая возможности повышения удобства взаимодействия с компьютером, сокращения финансовых и временных затраты при выполнении многих прикладных задач. Современные диалоговые системы являются виртуальными помощниками в сферах медицинского сопровождения, банковского обслуживания, обучения иностранным языкам. Большое распространение такие приложения получают благодаря реализации в них речевого интерфейса.

Взаимодействующие с пользователем на естественном языке интеллектуальные диалоговые системы предоставляют услуги заказа и бронирования авиабилетов. Использование клиентами такого сервиса повышает удобство взаимодействия с информационными ресурсами, позволяет существенно сократить время на выполнение требуемого запроса.

Сегодня с помощью диалоговых систем можно не только забронировать авиарейс, вызвать такси, выбрать ресторан, но и повысить производительность обслуживания потока клиентов, предоставить удобный сервис взаимодействия для информирования о конкретном реализуемом продукте. Применение чат-бота с речевым интерфейсом, работающего на основе обработки запроса на естественном языке, помогает пользователю в поиске товаров и услуг. При этом уточнение характеристик товаров или услуг осуществляется в процессе диалога, что позволяет значительно увеличить скорость и эффективность поиска

Сегодня для оценивания эффективности диалоговых систем используются наборы многих самых различных показателей. Для оценки целеориентированных систем может быть использован «процент диалогов, в которых пользователь достиг своей цели». Другие системы не могут быть оценены с помощью этого показателя, потому что они предназначены для выдачи ответов, соответствующих контексту диалога. В таких случаях оценивание выдаваемых системой ответов может осуществляться с использованием показателя перплексии как обратной вероятности тестового набора, нормализованной по количеству слов. Однако такой подход не всегда позволяет оценить адекватность ответа.

Для оценки эффективности чат-бот-приложений предлагается использовать KPI-метрики, которые используются в бизнес-структурах как показатель деловой активности сотрудников, например, активность, индекс потребительской лояльности и т.п. Кроме того, предлагается использовать стандартные технические показатели программного обеспечения (показатель диалогового потока, показатель качества классификации сообщений, показатель качества извлечения данных из пользовательских сообщений и др.). В публикациях в числе индикаторов качества чат-ботов представлены показатели количества активных пользователей чат-бота и удовлетворенность пользователей его работой.

Предложим ряд других качественных показателей оценивания чат-ботов:

- соотношение заработной платы сотрудника и стоимости соответствующего чат-бота;

- степень замещения персонала чат-ботом;

- полнота выполнения чат-ботом требуемого функционала;

- распространенность чат-бота, успешное применение в отрасли;

- совместимость программного обеспечения чат-бота с используемыми смежными решениями;

- степень отказоустойчивости чат-бота;

- масштабируемость (перспектива расширения функционала);

- степень соответствия тенденциям развития информационных технологий.

Отметим следующие качественные характеристики виртуальных помощников:

- понимание произвольной речи и команд;

- способность решения нетипичных задач;

- простота установки приложения на требуемой платформе;

- приспособленность к настройке и корректировке работы.

В числе важных качественных показателей, характеризующих диалоговых агентов, называются «не только возможности выполнения конкретных повседневных сценариев пользователя», но и когерентность, т.е. возможность поддержания связного общения, а также консистентность как способность имитировать поведение одного и того же человека, выдавая ответы, согласованные друг с другом по смыслу.

При анализе систем разговорного диалога применяется также метрика для оценки взаимодействия между компьютером и человеком, называемая качеством взаимодействия. При этом для каждого типа диалога мера качества взаимодействия имеет разное значение.

В качестве критериев оценки поддерживаемых виртуальными помощниками диалогов используем максимизацию следующих показателей:

- объем словарного запаса;

- степень разнообразия тем диалогов;

- количество используемых виртуальным помощником эмоций.

Для оценивания ответов, выдаваемых диалоговой системой, могут использоваться стандартные показатели качества классификации (точность, полнота и F1-мера). Так, в статье Юсупова И.Ф. «Диалоговый агент с обучаемым диалоговым менеджером» представлены результаты исследования по оценке качества диалогового менеджера. При этом оценивались отдельно функции генерации вопросов, реферирования, проверки ответа, другие разговорные навыки. Однако авторы работы «Применение глубокого обучения к моделированию диалога на естественном языке» Е.В. Чистова, А.О. Шелманов, И.В. Смирнов утверждают, что применение классических критериев точности, полноты и F1-меры для оценивания моделей диалоговых систем не целесообразно, потому, что вариативность синтезируемых ими последовательностей слов довольно высока.

Метрики точности и полноты применяются для оценивания качества систем вопросно-ответного поиска. В рамках теории информационного поиска для оценивания показателя точности используется следующее выражение:

, (1.1)

где  – число найденных поисковой системой релевантных документов;  – число найденных поисковой системой нерелевантных документов.

Оценивание показателя полноты информационного поиска осуществляется с помощью следующего выражения:

, (1)

где  – число не найденных поисковой системой релевантных документов.

Величины, используемые для оценивания характеристик информационного поиска, обозначаются в соответствии с таблицей 1.

Таблица 1 – Обозначения величин, используемых для оценивания характеристик информационного поиска

|  |  |  |
| --- | --- | --- |
| Число документов | Релевантных | Не релевантных |
| Найденных |  |  |
| Не найденных |  |  |

Особенности оценивания характеристик информационного поиска иллюстрируются на рисунке 1.



Рисунок 1 – Величины для оценивания характеристик информационного поиска

На рисунке 1 серым цветом обозначено множество релевантных запросу документов, имеющихся на информационных ресурсах. Белым цветом окрашено множество не релевантных запросу документов, имеющихся на информационных ресурсах. Овал с пунктирной линией соответствует множеству документов, найденных поисковой системой.

Анализ показал, что применение метрик полноты и точности для оценивания характеристик диалоговых систем сопряжено со значительными сложностями. Они связаны с тем, что в рамках оценивания качества информационного поиска вычисление значений этих метрик основано на подсчете конечного числа найденных (не найденных) системой релевантных (нерелевантных) документов. Чтобы применить данные метрики для оценивания характеристик диалоговых систем, потребуется вычислить количество содержащейся в ответе требуемой пользователю информации или другой информации, что весьма проблематично.

В отдельных случаях для оценки эффективности чат-ботов можно использовать метрики, которые были разработаны для оценки качества моделей машинного перевода, например, BLEU и METEOR – показатели различий между компьютерным переводом и эталонным пользовательским переводом. Метрика BLEU, вычисляемая как процентная доля n-грамм, совпадающих в вариантах эталонного и машинного перевода, имеет ряд модификаций . По мнению ряда специалистов, метрики, применяемые для определения качества машинного перевода, не всегда пригодны для оценивания эффективности диалоговых систем. Данная позиция обосновывается тем, что, во-первых, эти метрики слабо коррелируют друг с другом, а во-вторых, «оценивают лишь лексическую гибкость и игнорируют отступление от контекста беседы».

В ряде случаев для оценки качества ответов диалоговых систем выбирается метрика Recalln@k, которая используется для анализа работы ранжирующих чат-ботов и показывает, какова доля релевантных ответов в числе *k* лучших. По мнению исследователей, «особенностью оценки качества диалога с чат-ботом является субъективность восприятия ответов пользователем, что затрудняет автоматизацию процесса оценки качества». Следуя этому тезису, справедливо считать допустимым оценивание релевантности ответов вручную, например, с использованием возможностей краудсорсинга.

Субъективное экспертное участие активно применяется при оценке эффективности различных программ обработки естественного языка. Помощь человека-эксперта, например, потребовалась на этапе «отладки алгоритма выявления знаний в HTML-коде электронных учебников и наполнения базы знаний диалоговой системы». Это позволило выявить следующие типы ошибок в работе программного модуля семантической сети: «некорректное восполнение недостающих членов предложения», «неразрешенные отсылки к предыдущему предложению, иногда к более ранним предложениям, выраженные местоимениями и другими частями речи», «некорректная обработка списков, которые могут представлять собой набор однотипных связей или связей различных типов», «некорректное выявление границ предложений генератором семантической сети, в частности из-за попадания точек внутрь тегов, рассматриваемых как единое целое», «предложения построены не по шаблону «подлежащее – сказуемое – вспомогательные члены предложения».

Интеллектуальные и коммуникативные возможности виртуального ассистента должны быть такими, что «человек, общающийся с ним, не сумеет точно установить, с кем он имеет дело, – с другим человеком или автоматом». Этот тезис соответствует представлениям А. Тьюринга, который предсказал возможность программирования работы машин так, что «шансы среднего человека установить присутствие машины через пять минут после того, как он начнет задавать вопросы, не поднимались бы выше 70%». Не случайно, тест Тьюринга был предложен для оценивания диалоговых систем. При этом предполагается использование «бинарного классификатора, играющего роль эксперта в тесте Тьюринга – он классифицирует поступающее высказывание как человеческое или машинное». Однако, по мнению ряда исследователей, такой способ оценки диалоговой системы не приемлем, т.к. он может служить лишь для «разоблачения машины», но не позволяет определить качество выполнения вопросно-ответных функций.

Наиболее естественным и распространенным подходом к оцениванию объектов, систем и их свойств является привлечение к этому процессу экспертов, специалистов в соответствующей предметной области. При этом, конечно же, каждый эксперт, будучи человеком, выдает субъективный результат. В целях повышения объективности получаемых результатов вполне оправдано стремление автоматизировать процесс оценивания, применив подходящий для этого научно-прикладной аппарат, в частности, технологии искусственного интеллекта. Способность человека рассуждать в неточных терминах, принимать решения в условиях неопределенности, успешно моделируется с помощью нечеткого логического вывода. Не случайно, имеется достаточное количество примеров применения этого аппарата для оценивания эффективности, качества различных процессов и систем, в том числе, и качества обработки естественно-языковых данных.

Следует принимать во внимание, что существенным недостатком систем с нечеткой логикой является сложность выбора параметров функций принадлежности входных и выходных переменных. Их настройка часто выполняется вручную на основе субъективных и недостаточно адекватных решений. Для устранения этого недостатка может быть рекомендована автоматическая нейросетевая настройка параметров нечеткого вывода, выполняемая с применением обучающих данных.

На основе представленных выше аргументов вполне обоснованной является попытка применить возможности нечеткого вывода и нейронных сетей для совершенствования процесса оценивания характеристик знание-ориентированных диалоговых систем.

**СПИСОК ЛИТЕРАТУРЫ**

1. Гостьева, О.В. Методика выбора IT-решений для автоматизации бизнес-процессов на предприятиях оборонно-промышленного комплекса / О.В. Гостьева // Известия Уральского государственного экономического университета. – 2015. – №5(61). – С. 101-109
2. Даутова, Л.Р. Обзор интеллектуальных диалоговых систем с интерфейсом на естественном языке / Л.Р. Даутова, А.Р. Нафикова // Colloquium-journal. – 2019. – № 1-5(25). – С. 24-25.
3. Игитян, Е.В. Анализ интеллектуальных диалоговых систем / Е.В. Игитян, И.К. Польщиков // Информационное обеспечение научно-технического прогресса: анализ проблем и поиск решений: Сборник статей Международной научно-практической конференции, Пермь, 25 июня 2022 года. – С. 46-48
4. Маннинг, К.Д. Введение в информационный поиск / К.Д. Маннинг, П. Рагхаван, Х. Шютце. – М.: Вильямс, 2011. – 528 с.
5. Модель интеллектуальной диалоговой системы для автоматизированного заказа билетов на основе семантического анализа / Ю. М. Монахов, Л. А. Артюшина, Е. О. Маков, М. Р. Исмаилова // Известия Института инженерной физики. – 2018. – № 3(49). – С. 77-81.
6. Низамов, А.А. Чат-боты для организации: возможности и реализация / А.А. Низамов // Интеллектуальный и научный потенциал XXI века: Материалы Международной (заочной) научно-практической конференции, Минск, 15 марта 2021 года. – Нефтекамск: Научно-издательский центр «Мир науки», 2021. – С. 41-44.
7. Отслеживание состояния целеориентированного диалога на основе БЕРТ / П.А. Гуляев, Е.А. Елистратова, В.П. Коновалов [и др.] // ТРУДЫ МФТИ. Труды Московского физико-технического института (национального исследовательского университета). – 2021. – Т. 13. – № 3(51). – С. 48-61.
8. Юсупов, И.Ф. Диалоговый агент с обучаемым диалоговым менеджером / И.Ф. Юсупов, Ю.М. Куратов // ТРУДЫ МФТИ. Труды Московского физико-технического института (национального исследовательского университета). – 2020. – Т. 12. – № 4(48). – С. 106-120
9. Lavie, A. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments / A. Lavie, A. Agarwal // Proceedings of the second workshop on statistical machine translation. – 2007. – Р. 228–231.
10. Igityan, E. Models and Algorithms for Evaluating the Dialogue System Efficiency / E. Igityan, K. Polshchykov // Collection of materials of the International Conference «Partial Differential Equations and Related Topics (PDERT’22)». – Belgorod, 2022. – P. 156–159.

**Игитян Елена Владимировна**

Белгородский государственный национальный исследовательский университет

Старший преподаватель кафедры информационных и робототехнических систем

Тел.: +7(4722) 24-54-13

E-mail: medevedeva\_e@bsu.edu.ru

**Польщиков Константин Александрович**

Белгородский государственный национальный исследовательский университет

Директор института инженерных и цифровых технологий

Тел.: +7(4722) 24-54-10

E-mail: polshchikov@bsu.edu.ru