НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2023 Том 23 № 1 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2023 Vol. 23 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2023-23-1-88-95
УДК 004.89

Диалоговая система на основе устных разговоров с доступом к неструктурированной базе знаний

Сергей Михайлович Маслюхин✉

ООО «ЦРТ-инновации», Санкт-Петербург, 194044, Российская Федерация
Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
maslyukhin@speechpro.com✉, https://orcid.org/0000-0002-9054-5252

Аннотация

Предмет исследования. Представлен подход к построению задачно-ориентированной диалоговой системы (разговорного агента) с доступом к неструктурированной базе знаний на основе устных разговоров с применением аугментации письменной речи, имитирующей результаты распознавания устной речи, комбинирования предсказаний классификаторов, генерации текста, дополненной поиском. **Метод.** Предложенный подход предусматривает аугментацию обучающих данных двумя способами: преобразованием текста в речь и обратно с помощью систем синтеза и распознавания речи; заменой части слов на основе матрицы спутываний системы распознавания речи. Диалоговая система с доступом к неструктурированной базе знаний решает задачу обнаружения высказывания, для которого необходим поиск дополнительной информации в неструктурированной базе знаний. С этой целью выполнено обучение моделей Support Vector Machine, Convolutional Neural Network, Bidirectional Encoder Representations from Transformers и Generative Pre-trained Transformer 2. Лучшие из представленных моделей использованы при формировании предсказания путем взвешенного комбинирования. Осуществлен выбор подходящего текстового фрагмента из базы знаний и генерация обоснованного ответа. Поставленные задачи решены путем адаптации модели генерации текста, аугментированной поиском Retrieval Augmented Generation. **Основные результаты.** Выполнена апробация подхода на данных конкурса 10th Dialogue System Technology Challenge (DSTC10). По всем метрикам, кроме Precision, новый подход значительно превзошел результаты базовых моделей, предложенных организаторами конкурса DSTC10. **Практическая значимость.** Результаты работы могут найти применение при создании чат-бот систем, обеспечивающих автоматическую обработку обращений пользователей на естественном языке на основе неструктурированной базы знаний, например базы ответов на часто задаваемые вопросы.

Ключевые слова

диалоговые системы, разговорные агенты, поиск информации, текстовая аугментация, генерация, аугментированная поиском

Благодарности

Исследование выполнено за счет гранта Российского научного фонда (№ 22-11-00128, https://rscf.ru/project/22-11-00128/).

Ссылка для цитирования: Маслюхин С.М. Диалоговая система на основе устных разговоров с доступом к неструктурированной базе знаний // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23, № 1. С. 88–95. doi: 10.17586/2226-1494-2023-23-1-88-95

Dialogue system based on spoken conversations with access to an unstructured knowledge base

Sergei M. Masliukhin✉

STC-innovations Limited, Saint Petersburg, 194044, Russian Federation
ITMO University, Saint Petersburg, 197101, Russian Federation
maslyukhin@speechpro.com✉, https://orcid.org/0000-0002-9054-5252

Abstract

This paper describes an approach for constructing a task-oriented dialog system (a conversational agent) with an unstructured knowledge access based on spoken conversations including: written speech augmentation that simulates

the speech recognition results; combination of classifiers; retrieval augmented text generation. The proposed approach provides the training data augmentation in two ways: by converting the original texts into sound waves by a text-to-speech model and then transforming back into texts by an automated speech recognition model; injecting artificially generated errors based on phonetic similarity. A dialogue system with access to the unstructured knowledge base solves the task of detecting a turn, which requires searching for additional information in an unstructured knowledge base. For this purpose, the Support Vector Machine, Convolutional Neural Network, Bidirectional Encoder Representations from Transformers, and Generative Pre-trained Transformer 2 models were trained. The best of the presented models are used in the weighted combination. Next, a suitable text fragment is selected from the knowledge base and a reasonable answer is generated. The tasks are solved by adapting the retrieval augmented text generation model Retrieval Augmented Generation. The proposed method tested on the data from the 10th Dialogue System Technology Challenge. In all metrics, except Precision, the new approach significantly outperformed the results of the basic models proposed by the organizers of the competition. The results of the work can be used to create chat-bot systems that provide automatic processing of user requests in natural language based on an unstructured knowledge access, such as a database of answers to frequently asked questions.

Keywords
dialogue systems, conversational agents, information retrieval, text augmentation, retrieval augmented generation

Acknowledgements
This research is financially supported by the Russian Science Foundation (No. 22-11-00128, https://rscf.ru/project/22-11-00128/).

For citation: Masliukhin S.M. Dialogue system based on spoken conversations with access to an unstructured knowledge base. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 1, pp. 88–95 (in Russian). doi: 10.17586/2226-1494-2023-23-1-88-95

Введение

Традиционно задачно-ориентированные диалоговые системы сконцентрированы на предоставлении информации и выполнении действий в соответствии с запросами пользователей, которые могут быть обработаны только с использованием программного интерфейса и обращения к базам данных. Однако в дополнение к запросам, ориентированным на задачи, у пользователей также существуют потребности, которые требуют большей информации, кроме предоставленной из внутренних баз данных. Например, большинство разговорных агентов могут помочь пользователям забронировать отель, ресторан или купить билеты в кино, но они не отвечают на дополнительные возникающие вопросы, например: имеется ли парковка транспортных средств; разрешено ли приводить в зарезервированное место домашних животных или детей; или какова политика отмены бронирования. Для обработки таких запросов обычно нет записи в базе данных. С другой стороны, соответствующая информация уже доступна на веб-страницах в виде раздела ответов на часто задаваемые вопросы и отзывов клиентов для многих из этих сценариев, выходящих за рамки возможностей программного интерфейса. Поскольку современные диалоговые системы не включают эти внешние источники информации в задачно-ориентированное моделирование диалога, пользователям необходимо самим посещать веб-сайты, чтобы узнать любую дополнительную информацию, выходящую за рамки программного интерфейса, что делает диалоговые взаимодействия неэффективными. Данное исследование направлено на поддержку сценариев, в которых диалог не прерывается, когда у пользователей есть запросы, которые выходят за рамки программного интерфейса, но потенциально необходимая информация доступна во внешних источниках.

В работах) [1–3] рассмотрены важность и сложность моделирования диалога с использованием внешних источников знаний в открытом домене. В настоящей работе подходы, используемые при создании разговорных агентов в открытом домене, адаптированы под задачно-ориентированное моделирование диалога. В отличие от открытого домена, который предполагает наличие и использование широкого набора знаний о мире, например из википедии, при задачно-ориентированном моделировании диалога система располагает ограниченным пулом знаний, необходимых системе для ответа на вопросы в рамках решаемой задачи. При этом большое значение имеет точность выбора знаний из неструктурированной базы и их правильное использование при генерации ответа. В процессе работы, в первую очередь, система выполняет задачу обнаружения — определяет необходимость обращения к базе знаний, в случае, когда ответ на запрос не может быть получен на основе программного интерфейса. Для решения этой задачи проведено обучение моделей: Support Vector Machine (SVM) [4], Convolutional Neural Network (CNN) [5], Bidirectional Encoder Representations from Transformers (BERT) [6] и Generative Pre-Trained Transformer 2 (GPT-2) [7] — лучшие из которых использованы при формировании предсказания путем взвешенного комбинирования. Далее системой осуществлена задача выбора — поиск в базе знаний текстового фрагмента, содержащего необходимую для ответа информацию. При решении данной задачи проведено сравнение популярных моделей поиска, таких как Dense Passage Retrieval (DPR) [8] и BiEncoder [9]. Последний этап — формирование обоснованного ответа, при котором решена задача генерации. Для этого проведено дообучение генеративной модели Bidirectional and Auto-Regressive Transformers (BART) [10]. Также рассмотрен подход генерации, дополненной поиском Retrieval Augmented Generation (RAG), решающий задачи выбора и генерации в рамках одной модели, обучающейся методом сквозного (end-to-end) обучения.

Другой важный аспект — подготовка системы к работе с устной речью. Применяемые в настоящее вре-

мя диалоговые системы достигли многообещающих результатов в письменных диалогах, однако их использование непосредственно в устной речи затруднено из-за различий в распределении данных, включающих несоответствие между письменной и устной речью, а также дополнительные шумы из-за ошибок систем распознавания речи. Рассмотрены два способа аугментации обучающих данных для адаптации системы к устной речи: преобразование текста в речь и обратно с помощью систем синтеза и распознавания речи; замена части слов в обучающих данных на слова из матрицы спутываний системы распознавания речи.

Постановка задачи

Для оценки результатов работы предложенного подхода использованы данные второй задачи второго трека международного конкурса 10th Dialogue System Technology Challenge (DSTC10) [11]. В отличие от первого трека конкурса 9th Dialogue System Technology Challenge (DSTC9) [12], система оценена не на письменных диалогах, а на устных. По условиям задачи кросс-доменные разговорные агенты дают ответы на вопросы, которые невозможно сгенерировать на основе программного интерфейса и записей в базе данных, поэтому им приходится извлекать связанные пары вопрос-ответ из неструктурированной базы знаний часто задаваемых вопросов. На основе полученных пар вопрос-ответ агенты генерируют ответ на естественном языке. Конкурс DSTC10 не предусматривал специального обучающего набора. Вместо этого использован обучающий датасет первого трека DSTC9, состоящий из 72 518 письменных диалогов. Этот набор представляет собой расширенную версию набора данных Multi-Domain Wizard-of-Oz dataset (MultiWOZ 2.1) [13–15], в который добавлены запросы, для ответа на которые необходимо обладать дополнительной информацией, содержащейся в текстовых фрагментах. Текстовые фрагменты собраны со страниц часто задаваемых вопросов и состоят из пар вопрос-ответ. Они относительно короткие и охватывают четыре разных домена: отель, ресторан, поезд и такси. Первые два разделены на сущности. На основе текстовых фрагментов сформирована неструктурированная база знаний. В то время как в валидационном наборе данных использованы аналогичные фрагменты и локация, что и в обучающем, в тестовом наборе представлена новая локация — Сан-Франциско и новые текстовые фрагменты, некоторые из которых относятся к новому домену — достопримечательности. Около половины из 4181 диалога тестового набора данных получены путем расширения набора данных MultiWOZ. Другая половина собрана из разговоров между людьми, посвященных туристическим поездкам в Сан-Франциско. Примерно десятая часть разговоров выполнена в устной форме и имеет аналогичные свойства, что и данные DSTC10. Отметим, что данные теста DSTC9 включают не результаты распознавания, а транскрипты, написанные человеком, без ошибок распознавания. Валидационный набор данных DSTC10 представляет собой те же 263 диалога, что и устная речь в тестовых данных DSTC9, но полученных с помощью системы распознавания речи. Тестовые данные состоят из 1988 дополнительных диалогов, собранных в той же локации. Сравнительные характеристики датасетов представлены в табл. 1.

В качестве системы автоматического распознавания речи при подготовке датасетов использована модель Wav2Vec 2.0 [16], предварительно обученная на 960 ч Librispeech [17] и дообученная с использованием 10 % данных из целевого домена. Далее получены топ-10 предсказаний с помощью языковой модели-декодера, построенной на основе KenLM [18], для всех письменных текстов из наборов данных MultiWOZ и DSTC9. Система достигла уровня ошибок в словах 24,09 %, что привело к заметным искажениям в данных. База знаний такая же, как и в тестовом наборе DSTC9. Дополнительная информация о датасетах представлена в [11] с описанием конкурса.

Описание подхода

Предложенный подход включает в себя несколько этапов и решение ряда задач. Первый этап — подготовка данных, направленная на формирование необходимых для обучения и оценки данных, которые затем использованы при решении всех последующих задач. Следующий этап включает в себя непосредственно реализацию задачно-ориентированной диалоговой системы с доступом к неструктурированной базе данных, которая решает задачи: обнаружения высказывания, для которого необходим поиск дополнительной информации в неструктурированной базе знаний; выбора подходящего текстового фрагмента из базы знаний; генерации обоснованного ответа. Общая структура предлагаемого подхода представлена на рисунке.

Подробная информация по каждому этапу представлена ниже.

Задача аугментации данных. На этапе подготовки данных решена задача аугментации обучающих данных для адаптации системы к данным, получаемым в результате распознавания речи.

Таблица 1. Сравнительные характеристики датасетов DSTC9 и DSTC10
Table 1. DSTC9 and DSTC10 comparative characteristics of datasets

Датасет	Набор данных					
	Обучающий		Валидационный		Тестовый	
	Письменная речь	Устная речь	Письменная речь	Устная речь	Письменная речь	Устная речь
DSTC9	72 518	0	9663	0	3918	263
DSTC10	72 518	0	0	263	0	1988

Рисунок. Общая структура предлагаемого подхода
Figure. The general structure of the proposed approach

Популярный способ имитации устной речи — зашумление письменной речи путем последовательного преобразования текстовых данных в речь и затем обратного преобразования речи в текст [19, 20]. В связи с тем, что система синтеза речи генерирует простую для распознавания речь, в рамках эксперимента в аудиосигналы дополнительно добавлены шумы, имитирующие запись в городской среде. Кроме применения английской модели синтеза речи, выполнен эксперимент с моделью для русского языка, поскольку она генерирует речь с сильным русским акцентом, которая сложнее для распознавания системами автоматического распознавания речи (Automated Speech Recognition, ASR).

Другой способ аугментации данных основан на анализе ошибок систем ASR в представленных организаторами конкурса DSTC10 валидационных данных, которые содержат 10 лучших предсказаний системы распознавания речи для каждого высказывания в диалоге. На основе сопоставления гипотез системы распознавания речи составлена карта частых акустических спутываний. Далее в обучающих данных случайным образом заменены слова, для которых есть созвучные из карты спутываний. Таким образом, датасет был размножен и получил версии, содержащие одну, пять и десять копий оригинального датасета с разными искажениями, внесенными в них.

В рамках данной работы исследованы перечисленные способы аугментации и проанализировано влияние доли аугментированных данных в обучении на качество работы моделей.

Задача обнаружения. Обнаружение запросов, требующих обращения к базе знаний, сводится к задаче бинарной классификации: модель должна определить, требуется дополнительная информация из базы знаний или нет.

Проведем ряд экспериментов по сравнению различных моделей-классификаторов в рамках решаемой задачи: SVM, CNN, BERT и GPT-2. На вход моделей подадим данные текущего запроса без истории диалога, так как эти высказывания могут привести к неверному выбору класса, что было подтверждено в ходе экспериментов. В результате экспериментов выбран ряд моделей для участия в формировании финального предсказания. Веса моделей при комбинировании подобраны с помощью линейной регрессии на валидационных данных.

Задача выбора. Цель задачи выбора состоит в поиске наиболее подходящего документа из базы знаний для заданного диалога. Для каждого текстового фрагмента из базы знаний модель предсказывает, подходит он или нет, и выбирает документ с наивысшей оценкой релевантности. Выполним сравнение моделей BiEncoder и DPR. Они имеют схожую структуру и состоят из двух кодировщиков на основе предобученной модели BERT. Один кодировщик использован для кодирования контекста, а второй кодирует текстовые фрагменты, представляющие собой кандидатов. Для оценки схожести

применено значение скалярного произведения между вектором контекста и матрицей кандидатов. Разница в подходах, кроме незначительных архитектурных различий, заключается в способе негативного семплирования [21]. Для модели BiEncoder дистракторами являются верные ответы для других запросов из того же батча — части данных, обрабатываемой моделью за один шаг обучения. При обучении модели DPR выбраны некоторые количества сложных дистракторов из текстовых фрагментов для той же сущности/домена и простых дистракторов для других сущностей/доменов из базы. Для DPR доступна модель, предобученная на задаче выбора релевантных текстовых фрагментов на данных из википедии[1].

Задача генерации. Задача генерации предполагает поддержание естественного диалога в соответствии с контекстом диалога и извлеченным из базы знаний текстовым фрагментом.

Для решения этой задачи применена модель последовательность-в-последовательность с условной (на основе некоторого контекста) генерацией BART. Модель состоит из кодировщика и декодировщика. Кодировщик формирует векторное представление входной последовательности, содержащее контекст диалога и извлеченный на предыдущем этапе из базы знаний текстовый фрагмент. Декодировщик на основе полученного представления генерирует ответ на естественном языке. Модель обучена на истинно верных фрагментах из базы знаний, так как при ошибке на предыдущем этапе модель в любом случае не может сгенерировать верный ответ из-за недостатка информации.

Генерация, дополненная поиском. Модель RAG объединяет в себе модели DPR и BART и решает сразу обе задачи: выбора подходящих документов и генерации соответствующего ответа. Первые эксперименты проведены с оригинальной моделью RAG[2]. Однако эксперимент оказался неудачным, так как оба компонента модели на данных из нового домена изначально давали очень плохие предсказания, и модель не обучалась. В связи с этим модель DPR была заменена на обученную на целевом домене модель BiEncoder. Модель BART была также дообучена на целевых данных. Затем модель RAG обучена в режиме end-to-end. Таким образом, получен наилучший результат на валидационных данных для подзадач выбора и генерации.

Эксперименты и результаты

Проведены эксперименты по увеличению количества данных за счет аугментации на задаче обнаружения. Для оценки результатов использованы стандартные метрики бинарной классификации: accuracy, precision, recall и F-score[3]. Зашумление данных путем последовательного преобразования текстовых данных в речь и затем обратного преобразования речевых данных в текст не дало положительных результатов. Это связано с тем, что синтезированные данные оказались слишком простыми для системы распознавания речи и были распознаны с нулевой ошибкой, в том числе при наложении дополнительных шумов. При использовании русскоязычной модели синтеза речи получен обратный результат. Вывод системы распознавания оказался полностью отличным от оригинального текста.

Аугментация данных на основе карты акустических спутываний системы распознавания речи позволила заметно повысить качество на проверочных данных DSTC10. В ходе экспериментов проведено сравнение результатов при добавлении различного объема аугментированных данных. Каждая порция аугментированных данных содержит такое же количество данных, как и оригинальный датасет, так как получается путем полного прохода алгоритма аугментации по нему. В результате наилучший результат получен при добавлении десяти порций аугминтированных данных (табл. 2).

Наилучшие результаты на задаче обнаружения показала модель GPT-2, обученная на аугментированных данных (табл. 3). Принято считать, что модель BERT лучше подходит для задачи классификации текстов. Вероятно, лучшие результаты модели GPT-2 связаны с размерами моделей: GPT-2 более чем в 3 раза больше модели BERT по количеству обучаемых параметров. Для расчета финального предсказания на тестовых данных DSTC10 использовано комбинирование лучших моделей. Для всех возможных комбинаций классификаторов, представленных в табл. 3, подобраны веса предсказаний моделей, участвующие в комбинировании, при помощи линейной регрессии на проверочном наборе данных. Результаты наилучшей комбинации: CNN и GPT-2 с весами 0,33 и 0,72 соответственно — совпадают с предсказаниями лучшей модели. Такой результат может быть связан с малым количеством примеров в проверочной выборке — 104 примера. Также подбор весов моделей при помощи линейной регрессии возможно недостаточно эффективен, и в дальнейшем планируются эксперименты с другими способами агрегации, например путем обучения двухслойного перцептрона.

В задаче выбора проведено сравнение моделей BiEncoder и DPR. Особое внимание было уделено обработке сложных случаев, когда модели предлагалось выбрать верный ответ из набора похожих кандидатов из одного домена. В этом случае была собрана тестовая выборка для раздельной оценки модели на сложных и простых наборах кандидатов. Для оценки результатов использована метрика полноты ранжирования (recall). Полнота показывает долю случаев, когда верный ответ оказался в топ-k предсказаний модели и обозначена — $Rn@k$, где n — общее количество кандидатов, из которых выбирает модель, k — максимальная позиция верного ответа в предсказании модели. При этом предсказание модели считается верным[4]. Подход

[1] [Электронный ресурс]. Режим доступа: https://github.com/facebookresearch/DPR (дата обращения: 17.12.2022).

[2] [Электронный ресурс]. Режим доступа: https://huggingface.co/facebook/rag-sequence-base (дата обращения: 17.12.2022).

[3] [Электронный ресурс]. Режим доступа: https://scikit-learn.org/stable/modules/model_evaluation.html (дата обращения: 17.12.2022).

[4] [Электронный ресурс]. Режим доступа: https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval) (дата обращения: 17.12.2022).

Таблица 2. Результаты сравнения производительности модели GPT-2 на задаче обнаружения при добавлении различного объема аугментированных данных

Table 2. Comparison results of the GPT-2 model performance on the detection task when adding different amounts of augmented data

Данные в обучении	Количество порций аугментированных данных	Accuracy	Precision	Recall	F-score
Оригинальные DSTC9 (обучающие)	—	0,863	0,972	0,673	0,795
Оригинальные DSTC9 (все)	—	0,920	0,977	0,817	0,890
Аугментированные	1	0,932	**0,989**	0,837	0,906
	5	0,939	0,968	0,875	0,919
	10	**0,947**	0,959	**0,904**	**0,931**

Таблица 3. Результаты экспериментов с моделями классификаторов для задачи обнаружения

Table 3. Results of experiments with classifier models for the detection task

Модель	Accuracy	Precision	Recall	F-score
SVM	0,905	**0,976**	0,779	0,866
CNN	0,905	0,954	0,798	0,869
BERT	0,928	0,967	0,846	0,903
GPT-2	**0,947**	0,959	**0,904**	**0,931**
Комбинация	**0,947**	0,959	**0,904**	**0,931**

Таблица 4. Результаты сравнения моделей BiEncoder и DPR на задаче выбора

Table 4. Comparison results of BiEncoder and DPR models on the selection task

Модель	Коэффициент скорости обучения	Количество простых негативных примеров	Количество сложных негативных примеров	*R*10@1 для сложных негативных примеров	*R*100@1 для простых негативных примеров
BiEncoder	$1 \cdot 10^{-5}$	—	—	**0,79**	**0,96**
DPR	$1 \cdot 10^{-5}$	12	12	0,50	0,43
	$2 \cdot 10^{-6}$	12	12	0,54	0,52
	$2 \cdot 10^{-6}$	1	15	0,52	0,28

Таблица 5. Результаты экспериментов сравнения конвейера BiEncoder и BART с моделью RAG на задаче генерации

Table 5. Comparison results of the BiEncoder and BART pipeline with the RAG model on the generation task

Модель	*R*12039@1	*R*12039@5	BLEU-1	Meteor	Rouge-1
BiEncoder + BART	**0,625**	0,760	0,132	0,150	0,143
RAG	**0,625**	**0,769**	**0,139**	**0,155**	**0,189**

с использованием ответов из батча в качестве дистракторов оказался очень эффективным, и по результатам эксперимента модель BiEncoder значительно превзошла модель DPR на обоих наборах кандидатов (табл. 4). Для модели DPR применение одинакового количества сложных и простых дистракторов в обучении показало наилучшие результаты.

Модель RAG незначительно превосходит модель BiEncoder при выборе подходящих документов (табл. 5). В процессе обучения модель лучше учится обрабатывать сложные случаи, когда несколько кандидатов похожи, при этом верным является только один. Для оценки результатов экспериментов на задаче генерации использованы метрики, основанные на пересечении между верной и предсказанной последовательностями слов: BLEU-1, Meteor и Rouge-1, где 1 — размер *n*-граммы, используемый при определении совпадений[1]. Модель RAG превосходит модель BART на задаче генерации, так как позволяет эффективно учитывать топ-20 предсказаний модели поиска релевантных текстовых фрагментов, а не топ-1.

Финальный результат получен в два этапа. Вначале получены предсказания для задачи обнаружения с помощью комбинирования моделей классификаторов. Далее для запросов, требующих обращения к базе знаний, вызвана модель RAG, которая извлекла релевантные документы и сгенерировала ответы. По всем метрикам, кроме precision, предложенный подход

[1] [Электронный ресурс]. Режим доступа: https://blog.paperspace.com/automated-metrics-for-evaluating-generated-text/ (дата обращения: 17.12.2022).

Таблица 6. Результаты сравнения предложенного подхода с базовой моделью на тестовых данных DSTC10
Table 6. Comparison results of the proposed approach with the baseline on DSTC10 test data

Подходы для сравнения	Задача обнаружения			Задача выбора	Задача генерации		
	Precision	Recall	F-score	*R*12039@1	BLEU-1	Meteor	Rouge-1
Базовая модель	**0,897**	0,674	0,769	0,495	0,125	0,152	0,136
Предложенный подход	0,888	**0,890**	**0,889**	**0,572**	**0,145**	**0,158**	**0,178**

значительно превосходит результаты базовых моделей, предложенных организаторами конкурса DSTC10 (табл. 6). Выбор модели в сторону наивысшего значения recall, а не precision, сделан осознанно, так как он позволяет пропустить как можно меньше запросов, требующих дополнительной информации из базы знаний.

Заключение

Полученные результаты могут найти широкое применение при создании чат-бот систем, обеспечивающих автоматическую обработку обращений пользователей в различных сферах жизни. Например, в службах поддержки банков, медицинских и государственных учреждений и т. д., или в системах бронирования билетов, ресторанов, отелей и т. п. Предложенный подход также может быть использован в голосовых роботах, благодаря адаптации моделей к устной речи.

В качестве дальнейшего развития подхода предполагается доработка метода аугментации данных путем последовательного преобразования текстовых данных в речь и затем обратного преобразования речи в текст. Предполагается расширение возможностей поиска не только в базе ответов на часто задаваемые вопросы, но и в базе отзывов пользователей.

Литература

1. Moghe N., Arora S., Banerjee S., Khapra M.M. Towards exploiting background knowledge for building conversation systems // Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018. P. 2322–2332. https://doi.org/10.18653/v1/D18-1255
2. Dinan E., Roller S., Shuster K., Fan A., Auli M., Weston J. Wizard of wikipedia: Knowledge-powered conversational agents // arXiv. 2019. arXiv:1811.01241. https://doi.org/10.48550/arXiv.1811.01241
3. Zhou K., Prabhumoye S., Black A.W. A dataset for document grounded conversations // Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018. P. 708–713. https://doi.org/10.18653/v1/D18-1076
4. Hearst M., Dumais S., Osuna E., Platt J. Scholkopf B. Support vector machines // IEEE Intelligent Systems and their Applications. 1998. V. 13. N 4. P. 18–28. https://doi.org/10.1109/5254.708428
5. Johnson R., Zhang T. Convolutional neural networks for text categorization: Shallow word-level vs. deep character-level // ArXiv. 2016. arXiv:1609.00718. https://doi.org/10.48550/arXiv.1609.00718
6. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding // Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Vol. 1 (Long and Short Papers). 2019. P. 4171–4186. https://doi.org/10.18653/v1/N19-1423
7. Radford A., Narasimhan K., Salimans T., Sutskever I. Improving language understanding by generative pre-training: preprint. 2018.
8. Karpukhin V., Oğuz B., Min S., Lewis P., Wu L., Edunov S., Chen D., Yih W.-T. Dense passage retrieval for open-domain question answering // Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020. P. 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550
9. Humeau S., Shuster K., Lachaux M., Weston J. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring // arXiv. 2020. arXiv:1905.01969. https://doi.org/10.48550/arXiv.1905.01969
10. Lewis M., Liu Y., Goyal N., Ghazvininejad M., Mohamed A., Levy O., Stoyanov V., Zettlemoyer L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension // Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. P. 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703
11. Kim S., Liu Y., Jin D., Papangelis A., Hedayatnia B., Gopalakrishnan K., Hakkani-Tur D. DSTC10 Track Proposal:

References

1. Moghe N., Arora S., Banerjee S., Khapra M.M. Towards exploiting background knowledge for building conversation systems. *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2322–2332. https://doi.org/10.18653/v1/D18-1255
2. Dinan E., Roller S., Shuster K., Fan A., Auli M., Weston J. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv*, 2019, arXiv:1811.01241. https://doi.org/10.48550/arXiv.1811.01241
3. Zhou K., Prabhumoye S., Black A.W. A dataset for document grounded conversations. *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 708–713. https://doi.org/10.18653/v1/D18-1076
4. Hearst M., Dumais S., Osuna E., Platt J. Scholkopf B. Support vector machines. *IEEE Intelligent Systems and their Applications*, 1998, vol. 13, no. 4, pp. 18–28. https://doi.org/10.1109/5254.708428
5. Johnson R., Zhang T. Convolutional neural networks for text categorization: Shallow word-level vs. deep character-level. *ArXiv*, 2016, arXiv:1609.00718. https://doi.org/10.48550/arXiv.1609.00718
6. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Vol. 1 (Long and Short Papers)*, 2019, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423
7. Radford A., Narasimhan K., Salimans T., Sutskever I. *Improving language understanding by generative pre-training*. Preprint. 2018.
8. Karpukhin V., Oğuz B., Min S., Lewis P., Wu L., Edunov S., Chen D., Yih W.-T. Dense passage retrieval for open-domain question answering. *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550
9. Humeau S., Shuster K., Lachaux M., Weston J. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv*, 2020, arXiv:1905.01969. https://doi.org/10.48550/arXiv.1905.01969
10. Lewis M., Liu Y., Goyal N., Ghazvininejad M., Mohamed A., Levy O., Stoyanov V., Zettlemoyer L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703
11. Kim S., Liu Y., Jin D., Papangelis A., Hedayatnia B., Gopalakrishnan K., Hakkani-Tur D. *DSTC10 Track Proposal:*

Knowledge-grounded Task-oriented Dialogue Modeling on Spoken Conversations. 2021.
12. Kim S., Eric M., Gopalakrishnan K., Hedayatnia B., Liu Y. Hakkani-Tur D.Z. Beyond domain APIs: task-oriented conversational modeling with unstructured knowledge access // Proc. of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2020. P. 278–289.
13. Budzianowski P., Wen T.-H., Tseng B.-H., Casanueva I., Ultes S., Ramadan O., Gašić M. MultiWOZ - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling // Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018. P. 5016–5026. https://doi.org/10.18653/v1/D18-1547
14. Eric M., Goel R., Paul S., Sethi A., Agarwal S., Gao S., Kumar A., Goyal A., Ku P., Hakkani-Tür D. Multiwoz 2.1: Multi-domain dialogue state corrections and state tracking baselines // Proc. of the Twelfth Language Resources and Evaluation Conference. 2020. P. 422–428.
15. Zang X., Rastogi A., Sunkara S., Gupta R., Zhang J., Chen J. MultiWOZ 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines // Proc. of the 2nd Workshop on Natural Language Processing for Conversational AI. 2020. P. 109–117. https://doi.org/10.18653/v1/2020.nlp4convai-1.13
16. Baevski A., Zhou H., Mohamed A., Auli M. Wav2vec 2.0: a framework for self-supervised learning of speech representations // Proc. of the 34th International Conference on Neural Information Processing Systems (NIPS'20). 2020. P. 12449–12460.
17. Panayotov V., Chen G., Povey D., Khudanpur S., Librispeech: An ASR corpus based on public domain audio books // Proc. of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015. P. 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964
18. Heafield K. KenLM: Faster and Smaller language model queries // Proc. of the Sixth Workshop on Statistical Machine Translation. 2011. P. 187–197.
19. Gopalakrishnan K., Hedayatnia B., Wang L., Liu Y., Hakkani-Tür D. Are neural open-domain dialog systems robust to speech recognition errors in the dialog history? an empirical study // Proc. Interspeech 2020. 2020. P. 911–915. https://doi.org/10.21437/Interspeech.2020-1508
20. Wang L., Fazel-Zarandi M., Tiwari A., Matsoukas S., Polymenakos L. Data Augmentation for Training Dialog Models Robust to Speech Recognition Errors // Proc. of the 2nd Workshop on Natural Language Processing for Conversational AI. 2020. P. 63–70. https://doi.org/10.18653/v1/2020.nlp4convai-1.8
21. Xu L., Lian J., Zhao W.X., Gong M., Shou L., Jiang D., Xie X., Wen J. Negative sampling for contrastive representation learning: A review // ArXiv. 2022. arXiv:2206.00212. https://doi.org/10.48550/arXiv.2206.00212

Knowledge-grounded Task-oriented Dialogue Modeling on Spoken Conversations. 2021.
12. Kim S., Eric M., Gopalakrishnan K., Hedayatnia B., Liu Y. Hakkani-Tur D.Z. Beyond domain APIs: task-oriented conversational modeling with unstructured knowledge access. *Proc. of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2020, pp. 278–289.
13. Budzianowski P., Wen T.-H., Tseng B.-H., Casanueva I., Ultes S., Ramadan O., Gašić M. MultiWOZ - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 5016–5026. https://doi.org/10.18653/v1/D18-1547
14. Eric M., Goel R., Paul S., Sethi A., Agarwal S., Gao S., Kumar A., Goyal A., Ku P., Hakkani-Tür D. Multiwoz 2.1: Multi-domain dialogue state corrections and state tracking baselines. *Proc. of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 422–428.
15. Zang X., Rastogi A., Sunkara S., Gupta R., Zhang J., Chen J. MultiWOZ 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. *Proc. of the 2nd Workshop on Natural Language Processing for Conversational AI*, 2020, pp. 109–117. https://doi.org/10.18653/v1/2020.nlp4convai-1.13
16. Baevski A., Zhou H., Mohamed A., Auli M. Wav2vec 2.0: a framework for self-supervised learning of speech representations. *Proc. of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*, 2020, pp. 12449–12460.
17. Panayotov V., Chen G., Povey D., Khudanpur S., Librispeech: An ASR corpus based on public domain audio books. *Proc. of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964
18. Heafield K. KenLM: Faster and smaller language model queries. *Proc. of the Sixth Workshop on Statistical Machine Translation*, 2011, pp. 187–197.
19. Gopalakrishnan K., Hedayatnia B., Wang L., Liu Y., Hakkani-Tür D. Are neural open-domain dialog systems robust to speech recognition errors in the dialog history? an empirical study. *Proc. Interspeech 2020*, 2020, pp. 911–915. https://doi.org/10.21437/Interspeech.2020-1508
20. Wang L., Fazel-Zarandi M., Tiwari A., Matsoukas S., Polymenakos L. Data Augmentation for Training Dialog Models Robust to Speech Recognition Errors. *Proc. of the 2nd Workshop on Natural Language Processing for Conversational AI*, 2020, pp. 63–70. https://doi.org/10.18653/v1/2020.nlp4convai-1.8
21. Xu L., Lian J., Zhao W.X., Gong M., Shou L., Jiang D., Xie X., Wen J. Negative sampling for contrastive representation learning: A review. *ArXiv*, 2022, arXiv:2206.00212. https://doi.org/10.48550/arXiv.2206.00212

Автор

Маслюхин Сергей Михайлович — ведущий научный сотрудник, ООО «ЦРТ-инновации», Санкт-Петербург, 194044, Российская Федерация; инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0000-0002-9054-5252, maslyukhin@speechpro.com

Author

Sergei M. Masliukhin — Leading Researcher, STC-innovations Limited, Saint Petersburg, 194044, Russian Federation; Engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0000-0002-9054-5252, maslyukhin@speechpro.com

Статья поступила в редакцию 05.10.2022
Одобрена после рецензирования 01.12.2022
Принята к печати 15.01.2023

Received 05.10.2022
Approved after reviewing 01.12.2022
Accepted 15.01.2023