Создание чат-бота: обзор архитектур и векторных представлений текста

Чижик А.В., Жеребцова Ю.А.

***Аннотация* — В данной работе приведен краткий обзор архитектур современных разговорных агентов (чат-ботов), выделены основные преимущества и недостатки каждого подхода. Представлен сравнительный анализ актуальных на сегодняшний день методов векторизации текстовых данных. Предложены результаты эксперимента по созданию русскоязычного чат-бота ранжирующего типа: проанализированы проблемы открытых источников данных с диалогами на русском языке, описан алгоритм обработки собранных данных для реализации бота, продемонстрирована его работа, а также количественная оценка качества ответов пользователю. Опубликован итоговый набор данных и программный код.**

***Ключевые слова* — обработка естественного языка, компьютерная лингвистика, машинное обучение, диалоговые системы, интеллектуальные чат-боты, эмбеддинги слов, векторные представления текста.**

I. Введение

На сегодняшний день одним из стремительно развивающихся направлений научных исследований является создание разговорного интеллекта, способного поддерживать полноценный человеко-машинный диалог на произвольное количество тем.

Разговорные агенты (диалоговые агенты, диалоговые системы (ДС), чат-боты) – это компьютерные системы, с которыми пользователь взаимодействует на естественном языке (ЕЯ). Разговорные агенты стали базисом современных персональных помощников, которые помогают выполнять повседневные задачи. Среди самых популярных можно назвать Яндекс.Алису, Siri от Apple, Google Assistant, Microsoft Cortana и Amazon Alexa.

Современные диалоговые системы принято делить на целеориентированные (closed-domain) и виртуальные собеседники с открытым доменом (open-domain). Целеориентированные диалоговые системы предназначены для решения конкретных заранее определенных задач пользователя, а виртуальные собеседники («болталки») необходимы для вовлечения пользователя в использование продукта с помощью имитации естественного разговора с ним. Диалог всегда обусловлен контекстом, который по мере развития диалога изменяется, задавая логическое движение в известном для собеседников направлении. Если логические соединения потеряны, это вызывает у общающихся эмоциональное разочарование. Значит, основным стимулом активного взаимодействия с разговорным агентом можно назвать эмоциональное удовлетворение пользователя от диалога с ним. На сегодняшний день создание универсального интеллектуального диалогового агента, сочетающего в себе не только возможности выполнения конкретных повседневных сценариев пользователя, но и поддержание связности беседы (когерентность), выдачу ответов, согласованных между собой по смыслу (имитация поведения одной личности, консистентность), является интересной, перспективной и при этом очень сложной задачей, которую исследователям и инженерам еще предстоит решить.

В рамках сложившегося подхода к разработке разговорных агентов, можно выделить три основных блока выполняемых ими задач: понимание естественного языка, управление диалогом и синтез ответа пользователю. Ядром системы является процесс анализа фразы пользователя в модуле обработки ЕЯ [2], который преобразует реплику пользователя в ее некоторое векторное представление [12], предварительно, как правило, выполнив ряд шагов обработки текста: сегментация, токенизация, нормализация, синтаксический разбор, выделение именованных сущностей, разрешение анафоры и неоднозначности [16]. Полученное векторное представление затем используется внутренней моделью системы для последующей выдачи ответа пользователю. Данный цикл обработки текста лежит в основе работы любого диалогового агента, а его сложность зависит от конкретной цели его создания.

Целью данного исследования является анализ современных подходов к разработке разговорных агентов в задаче поддержания естественного диалога на примере эксперимента по созданию чат-бота, взаимодействующего с пользователем на русском языке. В рамках проведенного эксперимента мы рассмотрели процесс создания чат-бота, изучив основные проблемы; предложили базовую реализацию и ее улучшения. Также был сформирован набор данных (dataset), достаточный для обеспечения разнообразия ответов бота в беседе на

Статья получена 9 апреля 2020.

Жеребцова Юлия Андреевна, ведущий инженер, Национальный Центр Когнитивных разработок, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург, Россия (julia.zherebtsova@gmail.com)

Чижик Анна Владимировна, кандидат культурологии, ведущий инженер, Национальный Центр Когнитивных разработок, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург, Россия (afrancuzova@mail.ru)

Равный вклад в работу. Equal contribution to the work

узкопрофильную тему (в качестве примера выбрана тема пленочной фотографии), а также связные ответы на общие фразы пользователя (за основу взят набор данных диалогов из фильмов).

II. Подходы к обучению чат-ботов

Виртуальный собеседник ELIZA [28] – одна из первых попыток реализовать естественный человеко-машинный диалог с человека с программой. Она представляет собой чат-бот, основанный на большом количестве созданных вручную шаблонов и эвристических правилах. Такой подход требует огромных человеческих ресурсов, направленных на предсказание возможных ветвлений беседы, что сильно ограничивает множество ответов [27]. В попытках борьбы с такими ограничениями исследователи разработали новый взгляд на возможность выстроить человеко-машинный диалог, воспользовавшись набором данных (data-driven chatbots) и моделями машинного обучения. Общая идея заключается в создании чат-бота, обученного на большой коллекции текстов человеческих диалогов. По методам обучения чат-боты можно разделить на порождающие и ранжирующие.

Порождающие чат-боты отвечают на сообщения пользователя с помощью алгоритмов генерации текста, предсказывая каждое следующее слово в реплике. Так в работе [20] было предложено не определять для чат-бота сценарий всего диалога, а попытаться обучить систему отвечать на последнюю реплику пользователя, используя подходы из задач машинного перевода [5]. На сегодняшний день основой порождающего подхода являются рекуррентные sequence-to-sequence encoder-decoder архитектуры [24], представляющие собой многослойные LSTM и GRU нейронные сети, использующие механизм attention [25]. Для задачи создания разговорного агента данная архитектура впервые была применена в работе [26]. Отметим, что обучения порождающих архитектур зачастую приводит к проблеме ответов слишком общими фразами (например, «я не знаю», «хорошо») [23], а также к проблеме неконсистентности ответов (на одинаковые вопросы, но сформулированные по-разному, бот отвечает также по-разному).

Несмотря на перспективность использования порождающих моделей, они являются достаточно непредсказуемыми в поведении для использования в коммерческих продуктах [7], поэтому до сих пор наибольшую популярность имеют ранжирующие чат-боты, выбирающие реплику из заранее заготовленного набора ответов. Для пар «реплика-ответ» (single-turn conversation) или «контекст-ответ» (multi-turn conversation) из набора данных строятся векторные представления одной размерности (encoder-encoder), после чего наиболее возможные ответы ранжируются в соответствии со значениями некоторой функции уместности между векторами (чаще всего – скалярное произведение или косинусное расстояние). Данный подход, получивший популярность в задачах информационного поиска, впоследствии во многих работах адаптировался для создания диалоговых систем [18, 10]. Выбор ответа ранжирующих диалоговых агентов осуществляется среди заранее заготовленных, поэтому важное преимущество такого подхода заключается в возможности ограничивать выдачу грамматически некорректных и неприемлемых ответов, которые могут присутствовать в обучающем наборе данных. В рамках нашего эксперимента была выбрана реализация ранжирующего чат-бота.

III. Модели векторизации текстов

Наиболее популярные на сегодняшний день алгоритмы создания векторных представлений текста основываются на идеях дистрибутивной семантики [8]: слова, встречающиеся в аналогичных контекстах с подобной частотой, семантически близки. При этом соответствующие им сжатые векторные представления (embeddings), находятся близко друг к другу по косинусной мере в некотором векторном пространстве.

Одним из самых базовых методов представления текстового документа в виде вектора является статистическая мера TF-IDF [15], которая вычисляется как произведение частотности слов в тексте к обратной частотности слова в коллекции документов. Настоящую популярность модели векторизации текстов информации приобрели в 2013 году после публикации работы [17] о подходе, который известен как Word2Vec. Получающееся векторное представление отражает контекстную близость слов: слова, встречающиеся в тексте рядом с одинаковыми словами, имеют высокое косинусное сходство, а значит можно говорить о семантической близости.

В результате обучения Word2Vec-модели создается фиксированный словарь, для пополнения которого потребуется обучить модель заново. Решение проблемы отсутствующих слов было предложено в рамках модели fastText [11, 3], которая является модификацией Word2Vec и рассчитывает векторные представления частей слов, из которых уже составляется вектор целого слова. На сегодняшний день существует множество других моделей векторизации текстов, среди которых стоит отметить модель GloVe [21], которая комбинирует алгоритмы матричных разложений и Word2Vec.

Перечисленные выше методы векторизации называют статическими, они имеют следующее ограничение: такие модели не учитывают многозначность и контекстно-зависимую природу слов, то есть для одного слова, встречающегося в разных контекстах, будет один усредненный эмбеддинг. За последнее время были разработаны контекстуализированные (динамические) языковые модели, которые позволяют вычислять эмбеддинги для слова (или целого предложения) в зависимости от его контекста употребления. Одним из главных событий 2018 года в области обработки ЕЯ и машинного обучения стала модель BERT [4], приложения которой позволили улучшить известные на момент его появления решения многих задач обработки текстов и компьютерной лингвистики. Также самые последние разработки в направлении развития

контекстуализированных эмбеддингов включают такие передовые модели как ELMO [22], XLNet [29] и GPT-2 [19].

Одна из общих проблем при работе с текстами в случае чат-ботов – мультиязычность: люди достаточно часто используют в разговорном языке заимствованные слова или целые цитаты [9]. В случае ранжирующих архитектур, смешивание языков может влиять на подсчет семантической близости между векторами. Современный подход к обеспечению мультиязычности заключается в подготовке модели, которая способна обобщать различные языки в общем векторном пространстве, где векторы одинаковых предложений находились бы близко друг к другу вне зависимости от языка входной реплики. Перспективным методом, реализующим идею эту идею, является LASER [1], разработанный группой исследователей Facebook. Кроме того, модель LASER переводит целые предложения в их векторные представления, что может быть преимуществом при создании эмбеддингов для работы ранжирующих чат-ботов.

IV. Описание эксперимента: подход к сравнению моделей векторных представлений

A. Особенности данных

Для создания чат-ботов требуется достаточно большой набор текстовых данных, содержащий диалоги на те темы, которые он должен поддерживать. Наличие в данных дополнительной контекстной информации, например, уникального идентификатора автора реплики (или его имени, его возраста и пола), а также данных о диалоге (время реплик, факт ответа на другую реплику), позволяет улучшить качество ответов чат-бота.

Среди наиболее популярных открытых источников данных, где можно собрать датасеты, подходящие для создания разговорного чат-бота, выделим следующие:

- *Субтитры к фильмам и сериалам*. Они содержат множество повседневных диалогов на общие темы. Однако в них имеется и множество специфических фраз, которые могут быть неуместны (например, диалоги из фэнтези и исторических фильмов) и отсутствует явное отделение одного диалога от другого.
- *Сервис микроблогинга Twitter*. Среди плюсов текстов в Twitter можно отметить точную разбивку на диалоги, наличие дополнительных данных об авторах и адресате реплики. Однако зачастую пользователи ведут обсуждения вокруг некоторого мультимедийного контента, который в большей степени определяет тему. Диалоги же на повседневные темы при этом ведутся достаточно разными группами пользователей, что может потребовать дополнительного исследования данных профилей авторов, чтобы впоследствии бот сохранял консистентность ответов.
- *Публичные чаты мессенджеров* (например, Telegram или Slack) являются источником для пополнения датасета узкопрофильными темами. Обучая чат-бот на таких данных, следует помнить, что в случае отсутствия модерации чата администраторами, в текстах может оказаться значительное количество реплик, содержащих язык ненависти, политические высказывания, обсценную лексику.
- *Другие источники данных*. Кроме упомянутых выше источников, можно также выделить комментарии в социальных сетях, различные веб-форумы на сайтах, сценарии фильмов и транскрипты телепередач, а также тексты художественной литературы.

Как мы видим, с одной стороны имеется множество источников, где можно получить открытые наборы данных с диалогами на русском языке, подходящие для обучения разговорного чат-бота. С другой стороны, в зависимости от цели создания бота, этих данных, возможно, будет недостаточно, т.к. в открытом доступе может не оказаться необходимого количества диалогов на требуемую тему.

B. Сбор данных

В основе эксперимента, описываемого в нашей работе, лежит идея создания чат-бота, который мог бы выдавать связный (когерентный) ответ на заданную реплику, создавая таким образом у пользователя впечатление осмысленного диалога. При этом ожидается, что такое поведение бот будет демонстрировать как на общие пользовательские реплики, так и на реплики, касающиеся какой-либо заранее выбранной узкой тематики. В качестве источников данных для реализации чат-бота нами были выбраны два публичных Telegram-чата, в которых участники обсуждают узкопрофильную тему – плёночную (аналоговую) фотографию: «Пленка (чат)»[1] (369 участников) и «Пленка>чат»[2] (492 участника), а также открытый набор данных с субтитрами фильмов и сериалов OpenSubtitles[3]. После объединения текстов из трех источников была получена общая коллекция данных, состоящая из 358 545 записей со следующими полями: уникальный идентификатор реплики; идентификатор реплики, ответом на которую является данная реплика; имя автора; имя адресата и текст реплики.

C. Предобработка данных

Поскольку при взаимодействии с ранжирующим чат-ботом, пользователь вероятнее всего введет реплику, которой не окажется в заранее заготовленном наборе ответов, то релевантными ответами будут считаться те реплики, контекст которых в имеющемся наборе данных будет наиболее семантически близок к введенной реплике. В работах [30, 14] авторы предлагают множество различных способов определения контекста ответа-кандидата для чат-бота, имеющего ранжирующую архитектуру. В нашей работе под контекстом ответа-кандидата мы будем понимать цепочку сообщений, которые предшествуют выбранной реплике, при этом данная реплика является явным ответом на предпоследнюю фразу в цепочке, сама при этом являясь в ней последней.

[1] https://t.me/filmpublic
[2] https://t.me/plenkachat
[3] http://opus.nlpl.eu/OpenSubtitles.php

Для того чтобы иметь возможность осуществить поиск наиболее релевантного ответа по контексту реплики, исходный датасет с диалогами был преобразован к виду «Контекст-Ответ». Далее текстовые данные были обработаны в следующей последовательности: разбиение текстов реплик на токены; удаление спецсимволов, ссылок и пунктуации; удаление стоп-слов и лемматизация токенов. После заключительного этапа предобработки был получен обновленный датасет, состоящий из 134307 пар текстов вида «Контекст-Ответ».

Для получения векторного представления были обучены следующие модели: TF-IDF (данная модель была выбрана в качестве базовой), Word2Vec, FastText, LASER (использовалась предобученная модель для русского языка из библиотеки laserembeddings[4]).

При наличии различных вариантов векторных представлений общего набора данных диалогов, векторизация входящего запроса пользователя и имеющегося контекста из базы ответов может быть вычислена усреднением векторов слов этих моделей или предложений. В частности, для модели Word2Vec были использованы два способа усреднения векторов слов: простое усреднение (Averaged Word2Vec) и взвешенное усреднение Word2Vec по TF-IDF (TF-IDF-weighted W2V).

D. Оценка качества векторизации

В качестве метрики оценки качества ответов чат-бота при различных методах векторизации текстов была выбрана метрика Recalln@k (Rn@k), которую зачастую используют для оценки качества работы ранжирующих чат-ботов R@k [6]. Данная метрика показывает долю релевантных реплик среди k лучших, выбранных моделью.

Для подсчета выбранной метрики R10@k на основе имеющегося набора данных был создан специальный тестовый датасет, состоящий из 134307 записей, где каждой имеющейся реплике-контексту был сопоставлен список из 10 ответов, среди которых находится один корректный ответ и 9 других случайно выбранных ответов, не являющихся ответами на данную реплику.

E. Результаты эксперимента

В ходе эксперимента результат выдачи ответа на входную реплику из тестового датасета был оценен в соответствии с выбранной метрикой R10@k. Результаты представлены в табл. 1.

Таблица 1. Оценка качества ответов чат-бота в соответствии с метрикой Recall@k для различных моделей векторизации текста

Метрика / Модель	TF-IDF	Averaged W2V	FastText	TF-IDF-weighted W2V	LASER
R_{10}@ 1	0.229	0.186	0.179	0.212	0.195
R_{10}@ 2	0.277	0.289	0.283	0.318	0.308
R_{10}@ 5	0.328	0.544	0.543	0.564	0.577

Результаты в табл. 1 демонстрируют, что при расчете метрики R10@1 лучшим стал метод TF-IDF, при R10@2 – TF-IDF-weighted W2V, а при R10@5 – LASER. При этом методы TF-IDF-weighted W2V и LASER показывают примерно одинаковые результаты и являются лучшими относительно остальных. Отметим, что особенностью оценки качества диалога с чат-ботом является субъективность восприятия ответов пользователем, что затрудняет автоматизацию процесса оценки качества. В связи с этим до сих пор ручная оценка уместности ответов остается актуальным инструментом (например, с помощью краудсорсинга) [13]. Поэтому для методов TF-IDF-weighted W2V и LASER было также проведено предварительное внутреннее тестирование работы чат-бота, работающего на этих моделях, на предмет восприятия ответов пользователем, где модель LASER показала себя лучше. Один из примеров получающегося диалога приведен ниже:

User: привет!
Bot: привет, привет
User: как дела?
Bot: хорошо, как ваши?
User: Какую камеру посоветуешь?
Bot: Хас // сленговое название марки «Hasselblad»
User: а «Зенит» хороший фотоаппарат?
Bot: а с чего вопрос сразу такой?)
User: что лучше - чб или цвет?
Bot: думаю, все зависит от твоего настроения
User: на какую пленку лучше снимать?
Bot: снимал на ильфорд, сейчас отснял кодак TX, и одно и то же, значит не в пленке дело
User: знаешь фильмы о фотографии?
Bot: Да, я нашел хорошие примеры только зарубежом
User: твой любимый фильм?
Bot: супер 8
User: серьезно?
Bot: серьёзно

V. Заключение

Важным направлением развития чат-ботов на сегодняшний день является попытка создания разговорного агента, способного поддерживать связный диалог. Популярная гипотеза состоит в том, что пользователь тем чаще будет возвращаться к взаимодействию с ботом, чем диалог с ним будет более непринужденным и похожим на естественный. Чат-боты ранжирующего типа наиболее представлены среди коммерческих реализаций, так как они быстро обучаются, они менее чем порождающие модели, подвержены проблеме слишком общих реплик, а также позволяют лучше контролировать выдачу неприемлемых ответов пользователю. В рамках проведенного эксперимента мы изучили основные проблемы чат-ботов ранжирующего типа на примере создания бота, отвечающего на узкопрофильную тему о пленочной фотографии, предложили базовую реализацию и ее улучшения. Были проанализированы особенности открытых источников данных с диалогами на русском языке, доступных на сегодняшний день, собран и проанализирован необходимый набор данных для обучения чат-бота, продемонстрирована его работа, а также количественная оценка качества ответов пользователю. Данные и код эксперимента опубликованы по ссылке[5].

[4] https://pypi.org/project/laserembeddings/

[5] https://github.com/yuliazherebtsova/plenka-chatbot

Библиография

[1] Artetxe M., Schwenk H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. CoRR. arXiv:1812.10464. 2018.

[2] Bellegarda J.R. Large-Scale Personal Assistant Technology Deployment: the Siri Experience. INTERSPEECH. 2013. P. 2029-2033.

[3] Bojanowski P., Grave E., Joulin A., Mikolov T. Enriching Word Vectors with Subword Information. arXiv:1607.04606. 2017.

[4] Che W., Liu Y., Wang Y., Zheng B., Liu T. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. CoRR. arXiv:1807.03121. 2018.

[5] Cho K. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. EMNLP . 2014. 1724-1734 pp.

[6] Dariu J., Rodrigo A., Otegi A., Echegoyen G., Rosset S., Agirre E., Cieliebak M. Survey on Evaluation Methods for Dialogue Systems. arXiv:1905.04071. 2019.

[7] Gao J., Galley M., Li L. Neural Approaches to Conversational AI. arXiv:1809.08267. 2019. 95 pp.

[8] Harris Z.S. Distributional structure. Word. 10. Issue 2-3. 1954. P. 146–162.

[9] Holger S., Douze M. Learning Joint Multilingual Sentence Representations with Neural Machine Translation, ACL workshop on Representation Learning for NLP. arXiv:1704.04154. 2017.

[10] Ihaba M., Takahashi K. Neural Utterance Ranking Model for Conversational Dialogue Systems. Proceedings of the SIGDIAL 2016 Conference. Association for Computational Linguistics. 2016. 393-403 pp.

[11] Joulin A., Grave E., Bojanowski P., Mikolov T. Bag of Tricks for Efficient Text Classification. arXiv:1607.01759. 2016.

[12] Jurafsky D., Martin J. H. Title Speech and Language Processing. 2nd edition. Prentice Hall. 2008. 988 pp.

[13] Liu C.-W., Lowe R., Serban I. V., Noseworthy M., Charlin L., Pineau J. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. arXiv:1603.08023. 2016.

[14] Ma W., Cui Y., Shao N., He S., Zhang W.-N., Liu T., Wang S., Hu G. TripleNet: Triple Attention Network for Multi-Turn Response Selection in Retrieval-based Chatbots. arXiv:1909.10666. 2019.

[15] Manning C. D., Raghavan P., Schütze H. An Introduction to Information Retrieval. Stanford NLP Group, Cambridge University Press. URL: https://goo.su/0LzL (дата обращения: 17.02.2020).

[16] Masche J., Le N.-T. A Review of Technologies for Conversational Systems Conference Paper in Advances in Intelligent Systems and Computing. 2018. P. 212-225.

[17] Mikolov T. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of Workshop at ICLR. URL: https://goo.su/0LZi (дата обращения: 16.03.2020).

[18] Nio L., Sakti, S., Neubig G., Toda T. Developing Non-goal Dialog System Based on Examples of Drama Television. Natural Interaction with Robots, Knowbots and Smartphones. 2014. P. 355-361.

[19] Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I. Language Models are Unsupervised Multitask Learners. Technical Report OpenAi. URL: https://goo.su/0LzI (дата обращения: 16.03.2020)

[20] Ritter A. Data-Driven Response Generation in Social Media. Conference on Empirical Methods in Natural Language Processing. Edinburgh. 2011. P.583-593.

[21] Pennington J., Socher R., Manning C. D. GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics. 2014. 1532-1543 pp.

[22] Peters M.E., Neumann M., Iyyer M. Deep contextualized word representations. arXiv preprint arXiv: 1802.05365. 2018.

[23] Sountsov P., Sarawagi S. Length bias in Encoder Decoder Models and a Case for Global Conditioning. arXiv:1606.03402. 2016.

[24] Sutskiever I., Vinyals O., Le, Q. V. Sequence to Sequence Learning with Neural Networks. arXiv:1409.3215. 2014.

[25] Vaswani A. Attention Is All You Need. arXiv: 1706.03762. 2017.

[26] Vinyals O., Le Q.V. A neural conversational model. arXiv preprint arXiv:1506.05869. 2015.

[27] Wallace R. The Elements of AIML Style. ALICE A.I Foundation, 2003. 86 pp.

[28] Weizenbaum J. ELIZA – A computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 1966. P. 36-45.

[29] Yang Z., Dai Z., Yang Y., Carbonell J., Salakhutdinov R., Le Q. V. XLNet: Generalized Autoregressive Pretraining for Language Understanding.arXiv:1906.08237. 2019.

[30] Zhang R., Lee H., Polymenakos L., Radev D. Addressee and Response Selection in Multi-Party Conversations with Speaker Interaction RNNs. arXiv:1709.04005. 2017.

Building a Chatbot: Architecture Models and Text Vectorization Methods

Anna V. Chizhik, Yulia A. Zherebtsova

***Abstract* — In this paper, we review the recent progress in developing intelligent conversational agents (or chatbots), its current architectures (rule-based, retrieval based and generative-based models) and discuss the main advantages and disadvantages of the approaches. Additionally, we conduct a comparative analysis of state-of-the-art text data vectorization methods which we apply in implementation of a retrieval-based chatbot as an experiment. The results of the experiment are presented as a quality of the chatbot responses selection using various R10@k measures. We also focus on the features of open data sources providing dialogs in Russian. Both the final dataset and program code are published. The authors also discuss the issues of assessing the quality of chatbots response selection, in particular, emphasizing the importance of choosing the proper evaluation method.**

***Keywords* — natural language processing, natural language understanding, dialogue systems, intelligent chatbot, retrieval-based chatbot, word embeddings, text vectorization.**

REFERENCES

[1] Artetxe M., Schwenk H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. CoRR. arXiv:1812.10464. 2018.
[2] Bellegarda J.R. Large-Scale Personal Assistant Technology Deployment: the Siri Experience. INTERSPEECH. 2013. P. 2029-2033.
[3] Bojanowski P., Grave E., Joulin A., Mikolov T. Enriching Word Vectors with Subword Information. arXiv:1607.04606. 2017.
[4] Che W., Liu Y., Wang Y., Zheng B., Liu T. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. CoRR. arXiv:1807.03121. 2018.
[5] Cho K. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. EMNLP . 2014. 1724-1734 pp.
[6] Dariu J., Rodrigo A., Otegi A., Echegoyen G., Rosset S., Agirre E., Cieliebak M. Survey on Evaluation Methods for Dialogue Systems. arXiv:1905.04071. 2019.
[7] Gao J., Galley M., Li L. Neural Approaches to Conversational AI. arXiv:1809.08267. 2019. 95 pp.
[8] Harris Z.S. Distributional structure. Word. 10. Issue 2-3. 1954. P. 146–162.
[9] Holger S., Douze M. Learning Joint Multilingual Sentence Representations with Neural Machine Translation, ACL workshop on Representation Learning for NLP. arXiv:1704.04154. 2017.
[10] Ihaba M., Takahashi K. Neural Utterance Ranking Model for Conversational Dialogue Systems. Proceedings of the SIGDIAL 2016 Conference. Association for Computational Linguistics. 2016. 393-403 pp.
[11] Joulin A., Grave E., Bojanowski P., Mikolov T. Bag of Tricks for Efficient Text Classification. arXiv:1607.01759. 2016.
[12] Jurafsky D., Martin J. H. Title Speech and Language Processing. 2nd edition. Prentice Hall. 2008. 988 pp.
[13] Liu C.-W., Lowe R., Serban I. V., Noseworthy M., Charlin L., Pineau J. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. arXiv:1603.08023. 2016.
[14] Ma W., Cui Y., Shao N., He S., Zhang W.-N., Liu T., Wang S., Hu G. TripleNet: Triple Attention Network for Multi-Turn Response Selection in Retrieval-based Chatbots. arXiv:1909.10666. 2019.
[15] Manning C. D., Raghavan P., Schütze H. An Introduction to Information Retrieval. Stanford NLP Group, Cambridge University Press. URL: https://goo.su/0LzL (дата обращения: 17.02.2020).
[16] Masche J., Le N.-T. A Review of Technologies for Conversational Systems Conference Paper in Advances in Intelligent Systems and Computing. 2018. P. 212-225.
[17] Mikolov T. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of Workshop at ICLR. URL: https://goo.su/0LZi (дата обращения: 16.03.2020).
[18] Nio L., Sakti, S., Neubig G., Toda T. Developing Non-goal Dialog System Based on Examples of Drama Television. Natural Interaction with Robots, Knowbots and Smartphones. 2014. P. 355-361.
[19] Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I. Language Models are Unsupervised Multitask Learners. Technical Report OpenAi. URL: https://goo.su/0LzI (дата обращения: 16.03.2020)
[20] Ritter A. Data-Driven Response Generation in Social Media. Conference on Empirical Methods in Natural Language Processing. Edinburgh. 2011. P.583-593.
[21] Pennington J., Socher R., Manning C. D. GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics. 2014. 1532-1543 pp.
[22] Peters M.E., Neumann M., Iyyer M. Deep contextualized word representations. arXiv preprint arXiv: 1802.05365. 2018.
[23] Sountsov P., Sarawagi S. Length bias in Encoder Decoder Models and a Case for Global Conditioning. arXiv:1606.03402. 2016.

[24] Sutskiever I., Vinyals O., Le, Q. V. Sequence to Sequence Learning with Neural Networks. arXiv:1409.3215. 2014.
[25] Vaswani A. Attention Is All You Need. arXiv: 1706.03762. 2017.
[26] Vinyals O., Le Q.V. A neural conversational model. arXiv preprint arXiv:1506.05869. 2015.
[27] Wallace R. The Elements of AIML Style. ALICE A.I Foundation, 2003. 86 pp.
[28] Weizenbaum J. ELIZA – A computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 1966. P. 36-45.
[29] Yang Z., Dai Z., Yang Y., Carbonell J., Salakhutdinov R., Le Q. V. XLNet: Generalized Autoregressive Pretraining for Language Understanding.arXiv:1906.08237. 2019.
[30] Zhang R., Lee H., Polymenakos L., Radev D. Addressee and Response Selection in Multi-Party Conversations with Speaker Interaction RNNs. arXiv:1709.04005. 2017.