Challenges of Building an Intelligent Chatbot

Anna Chizhik [0000-0002-4523-5167]* and Yulia Zherebtsova [0000-0003-4450-2566]*

ITMO University, St.-Petersburg, Kronverkskiy Prospekt, 49, 197101
afrancuzova@mail.ru, julia.zherebtsova@gmail.com

Abstract. There can be no doubt that the way of human-computer interaction has changed drastically over the last decade. Dialogue systems (or conversational agents) including voice control interfaces, personal digital assistants and chatbots are examples of industrial applications developed to interact with customers in a human-like way using natural language. With continued growth in messaging applications and increasing demand for machine-based communications, conversational chatbot is likely to play a large part in companies' customer experience strategy. As systems designed for personalized interaction with users, conversational chatbots are becoming increasingly sophisticated in an attempt to mimic human dialogue. However, building an intelligent chatbot is challenging as it requires spoken language understanding, dialogue context awareness and human-like aspects demonstration. In this paper, we present the results of data-driven chatbot implementation in order to better understand the challenges of building an intelligent agent capable of replying to users with coherent and engaging responses in conversation. Developed chatbot demonstrates the balance between domain-specific responding and users' need for a comprehensive dialogue experience. The retrieval-based model, which achieved the best dialogue performance, is proposed. Furthermore, we present the datasets collected for the purpose of this paper. In addition, natural language understanding issues and aspects of human-machine dialogue quality are discussed in detail. And finally, the further studies are described.

Keywords: Natural Language Processing, Dialogue Systems, Conversational AI, Intelligent Chatbot, Retrieval-Based Chatbot, Word Embeddings, Text Vectorization.

Introduction

Intelligent dialogue agents are designed to conduct a coherent and emotionally engaging conversation with users. Chatbots became a basis of modern personal assistants which help users to perform everyday tasks. Among the most popular are Apple Siri, Google Assistant, Microsoft Cortana, Amazon Alexa and Yandex.Alice.

There are two major types of dialogue systems: goal-oriented (closed-domain) and open domain (i.e., chatbots or chitchats). Goal-oriented dialog systems are primarily

[*] Equal contribution to the work

built to understand the user request within a finite number of pre-defined agent skills (e.g., play music or set a reminder). Chatbots are to involve users in some kind of intelligent conversation in order to improve their engaging experience.

Building an intelligent conversational agent interacting with people in a human-like way is an extremely challenging task complex task, meanwhile it is a perspective and promising research direction of the field dialogue systems [1, 14].

Modern dialogue system architecture includes three main modules: natural language processing (NLP), dialogue manager, and natural language generation (NLG). The core of a dialogue system is analysis of user utterance inputted in NLP module [5]. Typically, in this module, the utterance is mapped to text vector representation (i.e., embeddings) [17]. Then vector representations are then used by the internal model to provide a response to the user. Chatbot could be considered intelligent if its responses are coherent and meaningful to the user. This behavior is highly dependent on the chatbot architecture and text vectorization methods.

The goal of this paper is analysis of modern approaches to the development of chatbots which could provide the user with emotionally satisfying and meaningful responses. First, we describe the historical background of conversational agents and consider the main data-driven architectures; in particular, we focus on the retrieval-based approach. Next, we briefly review the state-of-the-art text vectorization models and present the results of comparative analysis. Then we describe our experiment of building a retrieval-based chatbot, starting with the process of train dataset collection that provides a wide range of chatbot about a specific topic. The topic of film/analogue photography has been chosen as an example. The basic implementation of chatbot and its improvements are proposed. Finally, the main challenges of building an intelligent conversational agent and future work are discussed.

1 Chatbot Architectures

Chatbots can be roughly divided into the following three categories based on the response generation architectures [4, 27]:

- rule-based chatbots, which analyze key characteristics of the input utterance and response to the user relying on a set of pre-defined hand-crafted templates;
- retrieval-based (IR-based) chatbots, which select response from a large pre-collected dataset and choose the best potential response from the top-k ranked candidates;
- generative-based chatbots, which produce a new text sequence as a response instead of selecting if from pre-defined set of candidates.

One of the most influential examples of conversational programs is ELIZA [42], the early dialogue system, which was designed at the MIT Artificial Intelligence Laboratory by Joseph Weizenbaum, simulated a human-like conversation as a psychologist. ELIZA is the rule-based chatbot that responds to the user combining complex heuristics and "if-then-else"-rules from the set of hand-crafted templates developed for the system specific domain. All early rule-based chatbots, including ELIZA, required much manual human effort and experts' knowledge to build, enhance and maintain such systems [41].

Thankfully, as a result of the recent progress in internet technology and data science, full data-driven architectures were proposed. Divided by machine learning approaches, there are two chatbot architectures using massive text collection analysis and natural language processing: generative-based and retrieval-based.

Generative-based chatbots reply to users applying natural language generation (NLG). They produce new responses from scratch word by word: given a previous conversation history, predict the most likely next utterance. The early response generative model proposed by Ritter in 2011 was inspired by Statistical Machine Translation (SMT) techniques [21]. Nowadays, the state-of-the-art in the NLG are Encoder-Decoder Sequence-to-Sequence (seq2seq) architectures [37] based on deep recurrent LSTM/GRU neural networks [7] with attention mechanism [33, 39]. The first adaptation of seq2seq architecture to the task of building a conversational agent was presented by [40]. Unquestionably, the fundamental advantage of generative-based chatbots is that they do not rely neither on a pre-defined set of rules nor on a responses repository. Thus, generative models tend to be more sustainable to new unseen input utterances and, as a result, to seem more coherent to the user. However, due to specificity of learning procedure, there are also some weaknesses of generative models: the problem of short informative responses (e.g. "I don't know", "okay") [35]; text generation grammatical and semantic mistakes that humans would never make; and dialogue inconsistency, where the model analyzes only the current user utterance without the previous context ("context-blindness"). The above mentioned problems are still unresolved despite attempts of researchers to handle them [18, 34].

Latest works [1] show researchers' high interest in generative-based chatbot architectures, thus rapid progress in this area can be expected. However, it is worth noting that generative models require a huge amount of training data and computational resources while they are still likely to respond unpredictably. Therefore, today, most of the industrial production solutions still remained retrieval-based [9].

In this paper we focused on the features of retrieval-based architecture. Retrieval-based chatbots do not generate new utterances but they select an appropriate grammatically correct response from a large set of pre-collected *Utterance-Response* pairs. Given a dialogue context, both input utterance and responses pairs are encoded into some vector space representation, then the system counting semantic similarity score for each pair (i.e. dot product or cosine similarity) selects the best response from high-matched candidates. This approach based on information retrieval paradigm [13] became quite popular in the area of conversational agents [12, 25, 26, 15]. Considering the learning process, there are two approaches for best response selection by retrieval-based model: supporting a single-turn conversation, matching current user utterance with candidate pairs without any context information, or conduct a multi-turn conversation, taking into account the previous utterances, which are typically defined as a dialogue context. Building a retrieval-based chatbot supporting a multi-turn conversation is a promising and challenging problem. In recent years, there has been growing interest in this research area [32, 45, 38].

In the next chapter we consider the concept of text similarity in detail and briefly review various vectorization models relevant for the task of retrieval-based chatbot implementation.

2 Text Vectorization Models

Text vectorization models that are popular today are based on the ideas of distribution semantics [10, 24]. According to the hypothesis of distributional semantics, words that occur in similar contexts with a similar frequency are considered semantically close. Corresponding dense vector representations which dimensions are much smaller than the dictionary's dimension (i.e., embeddings) are close to each other by the cosine measure in a word vector space.

One of the most basic vectorization methods is the statistical measure TF-IDF [22], it determines the word importance to the document in a text collection. The TF-IDF is the product of the frequency of words in the text and the inverse frequency of the word in the collection of documents. So the value of TF-IDF increases proportionally to the number of times a word appears in the document. TF-IDF vectors have size equal to the dictionary size, and it can turn out to be quite large. TF-IDF vectors will be close only for those documents which contain the matching words [2].

Text vectorization models gained a wide popularity in 2013 after Tomas Mikolov publication [23] on the approach known as Word2Vec. This approach has two implementations: CBOW (continuous bag of words) and Skip-Gram. CBOW model predicts the probability of each word of text in a particular context, while the Skip-Gram model calculates the probability of a context around a particular word. Word2Vec embeddings capture semantic similarity of words, that is semantically close words will have high cosine similarity in the model vector space.

However, extension of Word2Vec vector space with new word embedding requires retraining the model. The solution of the missing words problem was proposed in fastText model [16, 6]. This model is Word2Vec modification, which produces character n-gram embeddings. Also, it is worth mentioning GloVe model [30] proposed by Stanford NLP Group at Stanford University. GloVe combines ideas of matrix factorization and Word2Vec approach.

Text vector representations described above are commonly referred to as "static word embeddings". One of the problems of static models is polysemy. The same words in different contexts will have the same embedding. The recent progress in approaches to text vector representation is contextualized (dynamic) language model. Contextualized models calculate word embeddings depending on its context. Thus, released in late 2018 BERT model, which helped researchers to reach a new state-of-the-art in most NLP problems, became, undoubtedly, the key achievement of the last years in the field of NLP. With regard to other successful contextualized language models, ELMO [31], XLNet [43] and GPT-2 [28] are particularly to be noted.

People often use foreign words or whole phrases in the spoken language [11]. Thus, multilingualism could be one of the challenges in building chatbots. Contextualized models allow a multilingual format, but separately trained models should be required for each language. There is another approach to multilingualism, which transfers NLP models from one language to scores of others by preparing a model that is able to generalize different languages in a common vector space. Then the vectors of the same statement in any language will end up in the same neighborhood closely placed. Developed by a group of Facebook researchers LASER embedding model [3] is the promising method which implements this idea.

The model maps entire sentences into the vector space, and that is the advantage in creating embeddings for retrieval-based chatbots. In the next section, we describe the steps of the retrieval-based chatbot implementation and present the results of comparison between the considered text vector models applied for this task.

3 Experiments and Results

3.1 Data Sources

Regardless of chatbot architecture, it requires a large dataset of natural language dialogs for training. Such a dataset should include all topics that are supposed to be discussed with the bot. Additional meta-information about the dialogs (i.e. author name and age, message date and time, or response links) can improve chatbot responses. The most notable conversational open data sources for Russian are the following:

- Movies and TV Series Subtitles. Subtitles can be a source of general conversation topics. However, the movie genre introduces the main theme of dialogs, thus the collected dataset must be analyzed for peculiar vocabulary. Another subtitles drawback is the lack of clear separation between dialogues.
- Twitter. Twitter messages in threads contain information about authors and reply details and conversations have clear boundaries. But Twitter users tend to discuss multimedia content, which makes the dialogue lexically and semantically narrow.
- Public Group Chats (i.e. Telegram, Slack). Public chats can provide a rich source of dialogues on specific topics (programming, history, photography, etc). However, it is necessary to remember that poorly moderated public group messages likely contain hate speech, political statements and obscene language.
- Other Web Sources. There are many other sources of conversational data that could be used for chatbots training: social networks discussions, forum threads, movie transcripts, fiction (i.e. plays), etc.

Depending on a practical goal, several data sources can be used for training a retrieval-based chatbot, but it still may not be enough for supporting a coherent conversation. Here it is also worth paying attention to ethical issues and removing offensive utterances and obscene language from the data.

The key idea of our experiment is creating a chatbot that could seem intelligent enough, responding to the input utterance coherently, which could make a good impression on users. The bot should behave that way both within small talk and within some pre-selected narrow topics, which users are interested in. As a subject of a specific topic we decided to choose analogue/film photography. Two public Telegram chats[1] and open set of subtitles[2] have been chosen as the data sources. Thus, the overall text collection consists of 358,545 records with the following columns: message identifier, reply message identifier, author, addressee and utterance.

[1] https://t.me/filmpublic, https://t.me/plenkachat

[2] http://opus.nlpl.eu/OpenSubtitles.php

3.2 Data Preprocessing

When users interact with a retrieval-based chatbot, they usually input a phrase that does not appear in predefined responses word-to-word.

Therefore, relevant responses could be selected only by the semantic similarity between the user's input and conversation context of candidate utterances. In the area of retrieval-based chatbots, various methods for defining the context have been proposed in many research papers [36, 20, 44, 21]. Since the chat-specific conversational data (i.e., Telegram chats) contains information about the authors and *reply_to* links, our dataset can be splitted into many short conversations of the form such as *start_utterance->response->...->response->last_utterance*. Figure 1 demonstrates multi-turn conversations of the dataset. The structure is a directed graph, where each node corresponds to the utterance labeled by message identifier and each edge corresponds to the relationship "is reply to" between messages.

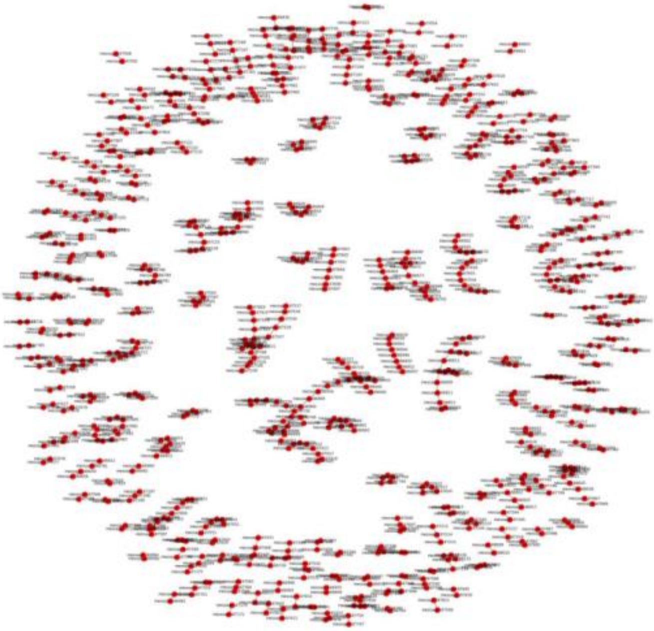

Fig. 1. Illustrated multi-turn conversations of training data

After multi-turns extraction, the initial dataset was transformed into the Context-Response form, where the Response is the last utterance of the turn and Context is all previous responses of that turn.

Further, the text data was pre-processed according to the following steps:

1. tokenization;
2. removal of special characters, links and punctuation;
3. removal of stop-words;
4. tokens normalization.

After the last step of data preprocessing, the final training dataset contained 134307 *Context-Response* pairs. The average *Context* length is 11 tokens and the average *Response* length is 9 tokens, which is, in fact, quite short for this kind of the retrieval-based task.

3.3 Results

Vector representation of text could be calculated by averaging its word embeddings. In particular, for the Word2Vec text vectors calculations two averaging word methods were used: simple averaging (Averaged Word2Vec) and weighted averaging W2V over TF-IDF (TF-IDF-weighted W2V).

For evaluation of chatbot responses based on the various text vectorization models, we use *Recall@k* metric. *Recall$_n$@k* (denoted *R$_n$@k* below) measures the percentage of relevant utterances among the top-*k* ranked *n* candidate responses [8]. This kind of metric is often applied to retrieval tasks and could be calculated automatically, but requires a validation set structured differently from training dataset. Concretely, we have created the dataset with 134307 records, where each record corresponds to three following columns: context, the ground truth response and the list of 9 false responses of training *Context-Response* pairs which have been chosen randomly. Thus, during the evaluation process, various *R$_{10}$@1, R$_{10}$@2 and R$_{10}$@5* measures have been calculated. Each model should select *1, 2* and *5* best responses among *10* possible candidates. Thus, the model's choice should be marked as correct if the ground truth utterance is ranked in top-*k*. Our experimental results are shown in Table 1.

It is worth noting that as a retrieval metric *R$_n$@k* has a significant drawback: in practice, there could exist more than one relevant response that could be marked as the ground truth. The appropriate responses thereby could be regarded as incorrect.

Table 1. Evaluation of chatbot performance based on various text vectorization models using R10@k measure

Model Metric	TF-IDF	Averaged W2V	FastText	TF-IDF-weighted W2V	LASER
R10 @ 1	0.229	0.186	0.179	0.212	0.195
R10 @ 2	0.277	0.289	0.283	0.318	0.308
R10 @ 5	0.328	0.544	0.543	0.564	0.577

According to Table 1, the different results for each text vectorization method have been demonstrated by the chatbot. For *R$_{10}$@1* the baseline TF-IDF has the highest score, for *R$_{10}$@2* - TF-IDF-weighted W2V and for *R$_{10}$@5* - LASER. TF-IDF-weighted W2V and LASER could be considered as the best overall models on the retrieval metrics. Even so, the model that performs well on the chosen retrieval metrics is not guaranteed to achieve good performance on a new response generation. Our assumption is that improvements on a model with regards to the *R$_n$@k* metric will eventually lead to improvements for the generation task. One the other hand, the human evaluation of conversational agents is still the most accurate and preferable approach [19]. Therefore, we evaluated the quality of two highly performed methods by human judgement (TF-IDF-weighted W2V and LASER). Finally, on the generation task, the chatbot based on LASER embeddings seemed significantly coherent, thus it has been considered as the best text vectorization model in our experiments.

Conclusions and Future Work

One of the most rapidly developing subfields of dialogue systems is an area of conversational agents (i.e. chatbots). Building an intelligent chatbot is a major issue of current business and research interests.

A strong product hypothesis is that the more conversational product interface is humanlike and intelligent, the more customers' digital experience is engaging and satisfactory. In this paper three main chatbot architectures have been briefly reviewed: rule-based approach and the fully data-driven retrieval-based and generative models. The advantages and disadvantages of the architectures have been also described. Nowadays, retrieval-based chatbots are the most commonly used conversational models which are built into business production solutions. Typically, retrieval-based models learn faster compared to generative models. They are less likely to have the problem of short general responses and more controllable for filtering grammatical mistakes and inappropriate language.

In this paper, the main challenges of data-driven conversational agents have been considered. We present the results of retrieval-based chatbot implementation, which keeps both a small talk conversation and conversation within a narrow topic of analogue photography in Russian. Semantic relations between context and potential responses are captured by text vector representation (word embeddings). It is a crucial technique for building a retrieval-based intelligent model of chatbot. In order to create a chatbot replying to users coherently and engagingly enough, the state-of-the-art text vectorization models have been compared and applied for our experiment. The LASER sentence embedding model has performed the best. The programming code and datasets have been shared in public repository[3].

Furthermore, we have analyzed current open web-sources of conversational data and outlined its main problems and features. It is essential to underline the critical need of high-quality dataset for training a retrieval-based chatbot. It is necessary to remember that poorly moderated conversational data likely contains offensive, toxic and noisy utterances, which must be removed from the dataset. This issue is one of the future research directions we plan to focus on.

References

1. Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., Le, Q. V.: Towards a Human-like Open-Domain Chatbot. arXiv:2001.09977. (2020).
2. Arora, S., Liang, Y., Ma, T.: A simple but tough-to-beat baseline for sentence, https://openreview.net/pdf?id=SyK00v5xx, last accessed 2020/02/17 (2017).
3. Artetxe, M., Schwenk, H.: Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. CoRR. arXiv:1812.10464. (2018).

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

4. Almansor, E., Hussain, F.K.: Survey on Intelligent Chatbots: State-of-the-Art and Future Research Directions. In: Complex, Intelligent, and Software Intensive Systems, P.534-543. (2020).
5. Bellegarda, J.R.: Large–Scale Personal Assistant Technology Deployment: the Siri Experience. INTERSPEECH. P. 2029-2033. (2013).
6. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. arXiv:1607.04606. (2017).
7. Cho, K.: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. EMNLP . 1724-1734 pp. (2014).
8. 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).
9. Gao, J., Galley, M., Li, L.: Neural Approaches to Conversational AI. arXiv:1809.08267. 95 pp. (2019).
10. Harris, Z.S.: Distributional structure. Word. 10. Issue 2-3. P. 146–162. (1954).
11. Holger S., Douze, M.: Learning Joint Multilingual Sentence Representations with Neural Machine Translation, ACL workshop on Representation Learning for NLP. arXiv:1704.04154. (2017).
12. Hu, B., Lu, Z., Li, H., Chen Q.: Convolutional neural network architectures for matching natural language sentences. In Advances in Neural Information Processing Systems. pp. 2042-2050 (2014).
13. Huang P.-S.: Learning Deep Structured Semantic Models for Web Search using Clickthrough Data, https://posenhuang.github.io/papers/cikm2013_DSSM_fullversion.pdf, last accessed 2020/02/17.. (2013).
14. Huang, M., Zhu, X., Gao, J.: Challenges in building intelligent open-domain dialog systems. arXiv preprint arXiv:1905.05709. (2019).
15. Ihaba, M., Takahashi, K.: Neural Utterance Ranking Model for Conversational Dialogue Systems. In: Proceedings of the SIGDIAL 2016 Conference. Association for Computational Linguistics. pp.393-403 (2016).
16. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of Tricks for Efficient Text Classification. arXiv:1607.01759. (2016).
17. Jurafsky, D., Martin, J. H.: Title Speech and Language Processing. 2nd edition. Prentice Hall. 988 p. (2008).
18. Li J., Galley, M., Brockett, C., Gao, J., Dolan, B.: A Diversity-Promoting Objective Function for Neural Conversation Models. arXiv:1510.03055. (2015).
19. 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).
20. Lowe, R., Pow, N., Serban, I. V., Pineau, J.: The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. arXiv:1506.08909. (2016).
21. 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).
22. Manning, C. D., Raghavan P., Schütze, H.: An Introduction to Information Retrieval. Stanford NLP Group, Cambridge University Press. URL: https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf (2009).
23. Mikolov, T. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of Workshop at ICLR, https://papers.nips.cc/paper/

5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf, last accessed 2020/02/17. (2013).

24. Osgood, C., Suci, G., Tannenbaum, P.: The measurement of meaning. University of Illinois Press. 354 p. (1957).
25. Nio, L., Sakti, S., Neubig, G., Toda, T.: Developing Non-goal Dialog System Based on Examples of Drama Television. In: Natural Interaction with Robots, Knowbots and Smartphones. P. 355-361. (2014).
26. Parakash A., Brockett, C., Agrawal, P.: Emulating Human Conversations using Convolutional Neural Network-based IR. arXiv:1606.07056. (2016).
27. Peng, Z., Ma, X..: A survey on construction and enhancement methods in service chatbots design. CCF Transactions on Pervasive Computing and Interaction. 10.1007/s42486-019-00012-3. (2019).
28. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language Models are Unsupervised Multitask Learners. Technical Report OpenAi, https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (2018).
29. Ritter, A.: Data-Driven Response Generation in Social Media. Conference on Empirical Methods in Natural Language Processing. Edinburgh. P. 583-593. (2011).
30. Pennington, J., Socher, R., Manning, C. D.: GloVe: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics. pp. 1532-1543 (2014).
31. Peters, M.E., Neumann, M., Iyyer, M.: Deep contextualized word representations. arXiv preprint arXiv: 1802.05365. (2018).
32. Serban, I., Lowe, R., Pow, N., , Pineau, J.: The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. arXiv preprint arXiv:1506.08909. (2015).
33. Shang, L., Lu, Z., Li, H.: Neural responding machine for short-text conversation. In Proc. of ACL-IJCNLP. pp. 1577-1586. (2015).
34. Shao, L., Gouws, S., Britz, D., Goldie, A., Strope, B., Kurzweil, R.: Generating High-Quality and Informative Conversation Responses with Sequence-to-Sequence Models. arXiv:1701.03185. (2016).
35. Sountsov, P., Sarawagi, S.: Length bias in Encoder Decoder Models and a Case for Global Conditioning. arXiv:1606.03402. (2016).
36. Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji,Y., Mitchell, M., Nie1, J.-Y., Gao, J., Dolan, B.: A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. arXiv:1506.06714. (2015).
37. Sutskiever, I., Vinyals, O., Le, Q. V.: Sequence to Sequence Learning with Neural Networks. arXiv:1409.3215. (2014).
38. Tao, C., Wu, W., Xu, C., Hu, W.: Multi-Representation Fusion Network for Multi-Turn Response Selection in Retrieval-Based Chatbots. In: ACM International Conference. pp. 429-437. (2019).
39. Vaswani A.: Attention Is All You Need. arXiv: 1706.03762. (2017).
40. Vinyals, O., Le, Q.V.: A neural conversational model. arXiv preprint arXiv:1506.05869. (2015).
41. Wallace, R.: The Elements of AIML Style. ALICE A.I Foundation, 86 pp. (2003).
42. Weizenbaum, J.: ELIZA – A computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 36–45. (1966).

43. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q. V.: XLNet: Generalized Autoregressive Pretraining for Language Understanding.arXiv:1906.08237. (2019).
44. Zhang, R., Lee, H., Polymenakos, L., Radev, D.: Addressee and Response Selection in Multi-Party Conversations with Speaker Interaction RNNs. arXiv:1709.04005. (2017).
45. Zhou, M., Wu, Y., Wu, W., Chen, X., Li, Z.: Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. ACL. arXiv:1612.01627. (2016).