eSIGNALS RESEARCH

Haaga-Helia

Article

Expert Work Automation in Healthcare: the Case of a Retrieval-Based Medical Chatbot

Lili Aunimo[1], Janne Kauttonen, Ari Alamäki

[1] Corresponding author

Abstract: This paper presents how medical expert work may be partially automated and made more interesting as input for routine conversation is handled by a software bot. However, the responsibility for treating the patient in the right way always stays with the medical doctor. The researchers describe how a retrieval-based one-to-one medical chatbot can be implemented for the Finnish language using neural networks based deep learning. The chatbot is evaluated using separate test data. The results show that a Top1 precision score of about 80% can be reached when the context size is 20. The Top1 precision score tells how often the chatbot ranks the correct answer as 1st among 10 candidates, where 1 answer is correct and 9 are wrong. The qualitative evaluation with healthcare services management shows that the healthcare industry shows great interest in medical chatbot systems. This is because they would both enhance the user experience and interestingness of work perceived by the medical doctors and at the same time make their work more productive. However, there is demand for practical systems integration and user interface development as well as for the development of task specific medical dialogue systems before medical chatbots become mainstream.

Keywords: chatbots, expert systems, deep learning, expert work automation, healthcare

Citation: Aunimo, L., Kauttonen, J., & Alamäki, A. (2022). Expert Work Automation in Healthcare: the Case of a Retrieval-Based Medical Chatbot. *eSignals Research, 3*(2). http://urn.fi/URN:NBN:fi-fe2022112466804

First submitted: 10.8.2022

Published: 24.11.2022

1. INTRODUCTION

This paper describes how medical expert work can be automated using a general-purpose medical chatbot. Chatbot is a software robot that participates in an online conversation that takes place in textual form. A general-purpose medical chatbot can address any medical issues whereas a special-purpose chatbot (also called a conversational agent) can only deal with a specific medical task. Chatbots are becoming more and more common. This is to a great extent because online live chat services operated by humans have been a popular way of communication between humans already for several decades and thus there is abundant data available for building machine learning models (Kucukyilmaz et al., 2008).

Chat services are in place in a variety of domains and tasks such as sales and customer support services in business, student tutoring in education and medical services in healthcare (Go & Sundar, 2019). Even though human mediated live chat services are commonplace in healthcare, to the best of our knowledge, there is not much research on chatbots that automate or assist a medical expert in delivering the live chat service to patients or to other healthcare professionals such as nurses. There is some research on health chatbot acceptability and usage from the point of view of patients (see e.g. Nadarzynski, et al. 2019). However, our research studies chatbots from the point of view of medical experts. The

Covid-19 pandemic increased chatbot adoption in healthcare (Amiri & Karahanna, 2022). However, our research does not study the automation of an entire conversation, but rather how a chatbot may assist the medical expert in delivering the live chat service. This research addresses the research gap by first exploring how such a chatbot can be built for a new language and secondly, it explores the suitability of a medical chatbot for expert work automation.

To address the research gap, this research describes the implementation and evaluation of a one-to-one general-purpose retrieval-based medical chatbot. In this context, a one-to-one chat means that only the patient or nurse and the physician are engaged in the conversation. A retrieval-based chatbot is based on a dataset of existing dialogues. A generative model would have been an alternative way of implementing a chatbot. In that model, the replies given by the bot are generated and not retrieved from a set of existing answers (Wu et al., 2019). The implementation described in this paper is based on proprietary medical live chat data in Finnish language.

The implemented medical chatbot is evaluated from a technical and a practical perspective. The technical evaluation is based on separate test data and standard metrics. It gives results comparable with previous research on retrieval-based chatbots. The practical evaluation is based on qualitative interviews with the management of a healthcare organization. It brings information on the level of automation achieved in a real-world setting, on user experience and on the benefits for the organization.

Research on general-purpose medical chatbots is important because they produce several benefits for the healthcare service providers. Firstly, people seek for health-related information from the internet, but it is often hard to find reliable information (Jacobs et al., 2017). Because of automation, a medical chatbot typically is a cost-effective way of offering reliable information when compared to a live chat operated by a medical expert. Secondly, a medical chatbot may take over the routine cases and thus increase the job satisfaction of a medical expert. Thirdly, the availability of medical services grows as the chatbot increases the productivity of medical expert work by automating the handling of some cases.

To the best of our knowledge, this is the first body of research work on automating medical chat in the Finnish language and on the consequences of automating medical chat from the point of view of the service provider. The part concerning the details of building the deep learning model are described in a research paper comparing several models by (Kauttonen & Aunimo, 2020). The contribution of the current study is twofold: Firstly, it adds knowledge to the debate on how to automate expert chat conversation for a new language. Secondly, it explores the consequences of expert work automation in a medical context.

2. AUTOMATION OF EXPERT WORK IN THE MEDICAL DOMAIN

Advanced medical chatbots aim to automate a part of healthcare practices and processes by replacing human conversation activities with computer-generated interactions. They may also utilize prior information stored in knowledge repositories and other databases. Thus, they potentially create value for the work of physicians, nurses and other healthcare professionals by providing new knowledge for decision-making. For example, IBM’s WatsonPaths can answer medical questions and suggest diagnosis or treatment (Lally et al., 2017).

Automating expert work by adopting AI in cognitively challenging tasks has become increasingly commonplace in the last years (Berente et al., 2021). However, expert work automation will replace significantly less entire jobs, professions, or occupations than single tasks (Chui et al., 2016).

Chatbots have been implemented in several different domains (Go & Sundar, 2019). Typically the term chatbot refers to text-based conversational systems that are open-domain, meaning that the chats may be on any topic, see e.g. (Zhou et al., 2018). Another term used for open-domain is general-purpose. However, in the medical domain, research on chatbots is scarce and the focus is on closed-domain conversational agents and dialogue systems. In this work, we use the terms dialogue system and conversational agent interchangeably.

In the medical domain, chatbots and dialogue systems have traditionally been mostly implemented in mental health. The first well-known conversational agent, Eliza, was programmed in 1966 to simulate a conversation with a psychotherapist (Weizenbaum, 1983). Abd-alrazaq et al., (2019) present an overview of the current use of chatbots in mental health. In a literature survey, they found out that chatbots are used in therapy, training and screening and that most of the implementations are rule-based.

Several researchers present surveys on the use of conversational agents in healthcare (see e.g. (Laranjo et al., 2018; Montenegro et al., 2019)). According to their studies, conversational agents are used in healthcare for the following tasks: training and practicing of skills among patients, education, prevention, educational assistance, diagnosis (including screening), self-monitoring and elderly assistance. The most important user groups of the agents are patients, physicians and students (Laranjo et al., 2018; Montenegro et al., 2019).

Conversational agents in the healthcare domain are mostly built using rule-based systems (Montenegro et al., 2019) or finite-state models (Laranjo et al., 2018). However, there is now days abundant data available for building chatbots using machine learning methods. Especially open-domain chatbots benefit from the data-driven approach because the manual modelling of all possibly upcoming conversational situations is practically impossible whereas a data-driven method can better cope with previously unseen input. The research presented in this paper shows how a modern data-driven and machine learning based chatbot may be implemented in the medical domain.

Data-driven and machine learning based methods for building chatbots may be categorized into two main groups: retrieval-based and those based on generative models. A retrieval-based chatbot uses a database of previously written utterances and merely prints as output the utterance with the highest confidence score in the specific situation (Wu et al., 2017) . The other commonly used way to implement a chatbot is to use a generative model which produces new utterances based on a natural language generation model (Wu et al., 2019). Both retrieval-based and generative chatbots have been implemented using a dataset of existing live chat dialogues between humans (Wu et al., 2017). Some chatbots only take into account the last utterance and search for an answer based on it, see e.g. (Wang et al., 2013). This type of chatbots are called single-turn chatbots. Other chatbots, like the one presented in this paper, take as input the preceeding context of the chat. The size of the preceeding context may vary. This type of chatbots are called multi-turn (Wu et al., 2017).

3. DATA AND METHODS

3.1. The Data

Our original dataset consisted of 29602 one-to-one chat dialogues between a patient and a healthcare professional or between two healthcare professionals. The data was obtained from a Finnish healthcare services provider and the main language of the data was Finnish. The data consisted of real chat dialogues from the years 2016 and 2017. Most of the dialogues were between a patient and a physician, while the rest were professional discussions between a nurse and a physician. The number of individual physicians was considerably smaller than the number of patients. However, this information was not available in the dialogue corpus. The dialogues were from the domain of general medicine. Each dialogue consisted of multiple separate dialogue turns (also called utterances) from the two speakers.

The original dataset contained highly sensitive information and it was anonymized before it was handed over to us, including any identifiers such as social security numbers, names, addresses and identities of speakers, including their roles. By roles we mean the expert, i.e., the physician or the non-expert, i.e., the patient or nurse. Our task was to choose the next physician response given a set of previous utterances. The set of previous utterances is called the context. The size of the context means the number of preceding utterances. In the following we describe the data pipeline and the methods used for building the model using machine learning techniques from the raw chat dialogue data. The data pipeline showing the transformations from raw data into the training, development and testing data sets is depicted in Figure 1.

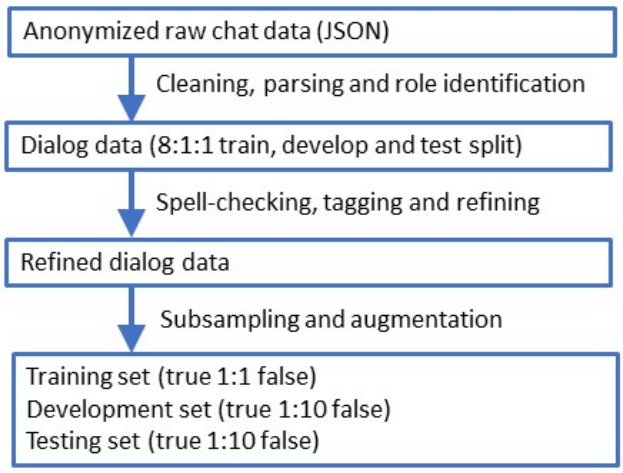

Figure 1: The process of transforming raw data into the training, development and testing data sets.

Preprocessing: Raw text data was preprocessed using the TurkuNLP pipeline (Zeman et al., 2018) based on deep neural networks (https://github.com/TurkuNLP/Turku-neural-parser-pipeline), which included tokenization and full morphological analysis including lemmatization and part-of-speech tagging, the Voikko spell-checker tool (https://github.com/voikko) and custom Python scripts. After running the texts through the TurkuNLP pipeline, processing steps were applied. The preprocessing contained standard steps such as removal of foreign language dialogs and lower-casing of all utterances.

Additionally, because the speaker roles were missing, we trained and applied a fastText binary classifier to add labels to utterances. Because the utterances between experts and customers were easily distinguishable and the model only needed to choose between two options, the model reached perfect accuracies in our tests. Further details of this are explained in Kauttonen and Aunimo (2020).

The aim of the preprocessing was to reduce noise and variability in the data by removing or transforming information that was not of interest for the task at hand. After this phase, the preprocessed dialogue corpus was split into three, non-overlapping parts (percentage in parenthesis): training (75%), development (12.5%) and testing (12.5%). The development set was used to monitor performance during training and stopping it after no improvement was detected. The best model was evaluated using the testing data set, as discussed later.

Data augmentation: For each context size, dialogues inside each data set (i.e., train, development and test) were subsampled in such a way that we created a sample from all available sub-partitions of the dialogueup to the maximum allowed context size. Here the minimum context length was set to 3. For example, if a dialogue had 10 turns and the maximum context size was 5, we generated 10 samples from it (i.e. 3 dialogues of length 5, 3 dialogues of length 4 and 4 dialogues of length 3). Note that the last utterance was always from a physician. As a result, as the maximum context size increases, so does the number of available samples from a dialogue. This approach was similar to what has been applied in augmenting image data (such as rotations and shifts, see, e.g. (Shorten & Khoshgoftaar, 2019)) and allowed us to utilize our dataset maximally. The downside is that, as we generated multiple samples from the same dialogue, the samples were no longer mutually independent inside the data partition.

3.2.Building the model

Model: We applied a model by Dong and Huan (Dong & Huang, 2018), who integrated character embeddings into Enhanced LSTM (Long Short Term Memory) method (ESIM) (Chen et al., 2017). In short, it is a deep neural network model based on LSTM cells and attention mechanism. The model was originally developed for Ubuntu Dialogue dataset in English and Douban Dialogue dataset in Chinese (Wu et al., 2017). The network contains two input heads; one for context and other for the response. We used the Python implementation of the model by Dong and Huang (https://github.com/jdongca2003/next_utterance_selection). The task to solve was binary classification with sigmoid cross-entropy loss function, which gives each sample (i.e., context text and response) a value between 0 (fully incompatible response) and 1 (fully matching response). This allows ranking of a set of responses for a given context. As the dataset was unlabeled and only contained positive samples (value 1), negative samples were generated by picking random physician utterance for a given context (value 0). Each correct response was matched by a randomly picked response, thus creating a fully balanced training data.

Parameters and training: We set limits for maximum utterance length to 80 tokens from the beginning, 3-25 turns in a dialogue and maximum 20 characters per word. For model parameters, we used LSTM cell count of 200, maximum word character count 20, word embedding dimension 200 and character embedding dimension 40, L2 regularization coefficient 1e-5 and batch size 64. No dropout was used. Word embeddings were initialized with standard word2vec

embeddings (skip-gram algorithm with windows size 5; (Mikolov et al., 2013)) which were trained using all sentences in training data partition. The ESIM model was trained as long as needed to reach the maximum performance for the development set, which was typically reached after 4-8 epochs. The model performance was measured via F1-score and Top1-score, which is a measure how often the model ranks the correct answer as 1st among 10 candidates, where 1 is correct and 9 are wrong (random). The higher these scores, the better the model. The random baselines for these scores were 10% for Top1-score and 0 for F1-score.

3.3.The Qualitative Interviews

The chatbot was presented to managers in a healthcare organization and an unstructured interview was performed to find out the expected benefits and drawbacks. All of the researchers participated as interviewers in the two interview sessions that were held, asked questions and took notes of the answers.

4. RESULTS

4.1. Model Evaluation

The number of individual dialogues was 24311 after preprocessing. Average turn count was 9 utterances (with 95th percentile 23) per dialogue. The best test accuracy was reached with context size 20 with F1-score 0.611 and Top1 precision 0.798 (80%). The model performed considerably better than the baseline, which was the accuracy of 0.10 (10%). The baseline accuracy is the accuracy that is reached by randomly selecting the response from the set of possible response candidates. In our test setting, we had 10 possible responses out of which 1 was correct and the other ones were false. The model was used to choose the correct answer among the 10 candidates.

The performance of our model (Top1 precision of 80%) was better than those previously obtained for related tasks (Top1 precision 76% for Ubuntu and 25% for Douban corpuses). However, as all three datasets are very different, no one-to-one comparison per se is possible.

4.2 Qualitative Interviews

In addition to testing the accuracy of the model with separate test data, the results provided by the medical chatbot were presented to two directors of the healthcare company for evaluating the validity of the modelling as well as the more general idea of automating experts' chat interaction. The medical chatbot received positive feedback. The directors believe that it would create value for the experts, and that the accuracy of the automatically generated response candidates was on an acceptable level.

To find out more precisely what are the benefits and possible obstacles of taking into use a medical chatbot, the researchers needed more information from the physicians who work with patients through the medical chat. However, the researchers were not able to test the model with physicians. Instead, they conducted an unstructured interview with the two directors of the healthcare company. These directors work in close cooperation with the physicians and with the IT department and thus they were able to provide insight about the value that the chatbot would bring to the physicians and also on the specific requirements of the physicians. The two most important aspects that arose from the interview were related to usability and to systems integration. Firstly, it turned out that the physicians working with the chat tools are experts in the usage of IT systems and

thus very demanding with regard to usability. They appreciate well-thought, logical and self-explanatory user interfaces. On the other hand, a medical chat system cannot work isolated from other IT systems such as patient records including a patient's diagnoses and previous prescriptions. Therefore, systems integration is an important part of the implementation.

5. DISCUSSION

The medical chatbot presented in this paper partly automated the work of the medical doctor in a chat conversation with a patient or a nurse. The chatbot presents the doctor an ordered list of potential answers given the preceding dialogue. More specifically, our task was to choose the most suitable response from a set of candidate responses to a given context, i.e. we needed to rank the responses. For this purpose, Lowe and colleagues (Lowe et al., 2015) created the well-known Ubuntu Dialogue Corpus with 930000 dialogues by scraping the Ubuntu operating system support channel in IRC. What differentiated the current task from the Ubuntu case, was the topic (medicine vs. information technology), language (Finnish vs. English) and roles (here only modeling the responses of physicians). Our dataset was also notably smaller (~3% of the Ubuntu Corpus).

The major limitations of the work were the relatively small dataset considering the large variability of vocabulary and topics in dialogs. Another major limitation of the work is the loss of information due to the anonymization process. This has been discussed in more detail in the work of Alamäki et al., 2019. Furthermore, many dialogues included external information not available from chat messages alone. For example, physicians often checked a database for patient-related laboratory results outside the chat. As a result, the context might not contain enough information to choose a proper response with a high confidence. Retrieval-based chatbots, such as the one developed here, can only respond with a predefined set of expert responses whereas generative models can create novel responses that could potentially match the context better. However, in healthcare domain generative models can be considered risky as the responses are not authored by actual healthcare professionals and there is little control of the generated output. A generative bot is probably better for a non-expert, chit-chat – type conversation.

One outcome of the evaluation discussions with the healthcare management is that chatbots are mostly useful for open-domain chat and dialogue systems for helping the medical doctor complete specific tasks such as screening for a specific illness. This is in line with previous research, see e.g. the work of (Wu et al., 2019) where they suggest that dialogue systems should be used instead of chatbots to accomplish specific tasks in various vertical domains such as a flight booking or buying an insurance. However, many cases are not as clear cut. For example, in the case of the medical chatbot, it is beneficial to have a medical domain-specific chatbot in the beginning of the conversation. After some utterances, a specific task, such as screening for depression or prescription of a specific drug may be detected. At that point, it may be beneficial to switch from a chatbot to a dialogue system.

The results of the qualitative unstructured interview indicate that a medical chatbot has high potential for bringing added value both to the interestingness of work as perceived by the physician and to the productivity of work. However, special attention should be paid both to the usability and to the integration of the chatbot to other IT systems of the healthcare company. In addition to studying the usability from the point of view of the doctor, also the acceptance of the new

technology from the point of view of the patients should be studied. Dobrowsky et al., 2021 present an overview of studies on how the interaction style of a chatbot affects its acceptance by users. All in all, we can state that the automation of routine utterances of a medical doctor in a chat is mostly already in place or coming soon. The mere existence of this research project shows that healthcare organizations have interest in exploring to opportunities offered by advanced chatbot technologies.

6. CONCLUSIONS

This paper presented how medical expert work may be partially automated and made more interesting as input for routine conversation is handled by a software bot. However, the responsibility for treating the patient in the right way stays with the medical doctor.

We described how a retrieval-based one-to-one medical chatbot can be implemented for the Finnish language using neural networks based deep learning. The chatbot was evaluated using separate test data. The results show that a Top1 precision score of 80% was reached. The Top1 precision score tells how often the chatbot ranks the correct answer as 1st among 10 candidates, where 1 answer is correct and 9 are wrong.

The qualitative evaluation with healthcare services management showed that the healthcare industry shows interest in advanced medical chatbot systems that would both enhance the user experience and interestingness of work perceived by the medical doctors and at the same time make their work more productive. However, there is still place for practical systems integration and user interface development as well as for the development of task specific medical dialogue systems before advanced medical chatbots become mainstream.

Acknowledgements

This work was supported by the BIG-research project funded by Business Finland and the AI-Driver -project funded by the Finnish Ministry of Education and Culture.

REFERENCES

Abd-alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., & Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, *132*, 103978. https://doi.org/10.1016/j.ijmedinf.2019.103978

Amiri, P., & Karahanna, E. (2022). Chatbot use cases in the Covid-19 public health response. *Journal of the American Medical Informatics Association*, *29*(5), 1000-1010.

Alamäki, A., Aunimo, L., Ketamo, H., & Parvinen, L. (2019). Interactive Machine Learning: Managing Information Richness in Highly Anonymized Conversation Data. In *IFIP Advances in Information and Communication Technology* (Vol. 568). https://doi.org/10.1007/978-3-030-28464-0_16

Berente, N., Gu, B., Recker, J., & Santanam, R. (2021). Managing Artificial Intelligence. *MIS Quarterly*, *45*(3), 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

Chen, Q., Zhu, X., Ling, Z.-H., Wei, S., Jiang, H., & Inkpen, D. (2017). Enhanced LSTM for Natural Language Inference. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1657–1668. https://doi.org/10.18653/v1/P17-1152

Chui, M., Manyika, J., & Miremadi, M. (2016). Where machines could replace humans-and where they can't (yet). *McKinsey Quarterly*.

Dobrowsky, D., Aunimo, L., Janous, G., Pezenka, I., & Weber, T. (2021). The Influence of Interactional Style on Affective Acceptance in Human-Chatbot Interaction--A Literature Review. *E-Signals Research*.

Dong, J., & Huang, J. (2018). Enhance word representation for out-of-vocabulary on Ubuntu dialogue corpus. *ArXiv Preprint ArXiv:1802.02614*.

Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, *97*, 304–316. https://doi.org/10.1016/j.chb.2019.01.020

Jacobs, W., Amuta, A. O., & Jeon, K. C. (2017). Health information seeking in the digital age: An analysis of health information seeking behavior among US adults. *Cogent Social Sciences*, *3*(1), 1302785. https://doi.org/10.1080/23311886.2017.1302785

Kauttonen, J., & Aunimo, L. (2020). Dialog Modelling Experiments with Finnish One-to-One Chat Data. In *Communications in Computer and Information Science: Vol. 1292 CCIS*. https://doi.org/10.1007/978-3-030-59082-6_3

Kucukyilmaz, T., Cambazoglu, B. B., Aykanat, C., & Can, F. (2008). Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing & Management*, *44*(4), 1448–1466. https://doi.org/10.1016/j.ipm.2007.12.009

Lally, A., Bagchi, S., Barborak, M. A., Buchanan, D. W., Chu-Carroll, J., Ferrucci, D. A., Glass, M. R., Kalyanpur, A., Mueller, E. T., Murdock, J. W., Patwardhan, S., & Prager, J. M. (2017). WatsonPaths: Scenario-Based Question Answering and Inference over Unstructured Information. *AI Magazine*, *38*(2), 59–76. https://doi.org/10.1609/aimag.v38i2.2715

Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A. Y. S., & Coiera, E. (2018). Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, *25*(9), 1248–1258. https://doi.org/10.1093/jamia/ocy072

Lowe, R., Pow, N., Serban, I., & Pineau, J. (2015). The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue* , 285–294.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, *26*.

Montenegro, J. L. Z., da Costa, C. A., & da Rosa Righi, R. (2019). Survey of conversational agents in health. *Expert Systems with Applications*, *129*, 56–67. https://doi.org/10.1016/j.eswa.2019.03.054

Nadarzynski, T., Miles, O., Cowie, A., & Ridge, D. (2019). Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *Digital health*, *5*, 1-12.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, *6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Wang, H., Lu, Z., Li, H., & Chen, E. (2013). A dataset for research on short-text conversations. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 935–945.

Weizenbaum, J. (1983). ELIZA — a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *26*(1), 23–28. https://doi.org/10.1145/357980.357991

Wu, Y., Wu, W., Xing, C., Xu, C., Li, Z., & Zhou, M. (2019). A Sequential Matching Framework for Multi-Turn Response Selection in Retrieval-Based Chatbots. *Computational Linguistics*, *45*(1), 163–197. https://doi.org/10.1162/coli_a_00345

Wu, Y., Wu, W., Xing, C., Zhou, M., & Li, Z. (2017). Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 496–505. https://doi.org/10.18653/v1/P17-1046

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., & Petrov, S. (2018). Turku Neural Parser Pipeline: An End-to-End System for the CoNLL 2018 Shared Task. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 1–21. https://doi.org/10.18653/v1/K18-2001

Zhou, H., Huang, M., Zhang, T., Zhu, X., & Liu, B. (2018). Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). https://doi.org/10.1609/aaai.v32i1.11325