Discover two Neural Machine Translation model variables' effects on Chatbot's performance

Youwei Wang*

Engineering, Virginia Polytechnic Institute, and State University, Blacksburg, VA 24061, United States

*Corresponding author's email: youwei@vt.edu

Abstract. To offer automatic online advice and help, chatbots are sophisticated conversational computer systems that resemble a human conversation. As chatbots' advantages grew, a variety of sectors began to use them extensively to give customers virtual support. Chatbots take advantage of methods and algorithms from two Artificial Intelligence areas: Machine Learning and Natural Language Processing. There are still several obstacles and restrictions to their use. In order to discover two NMT model variables' effects on chatbot performance, this paper does several experiments on a deep neural network chatbot model. Two straightforward and useful kinds of attentional mechanisms are used in this chatbot model: a local technique that only considers a small subset of source words at a time, as opposed to a global approach that always pays attention to all source words. This paper conducts experiments to examine how different model variables affect chatbot performance. This paper created a question template with eight general questions to test chatbot performance. Through the whole experiment results, increasing the number of iterations and increasing the dataset scale can improve the vocabulary and logic of the chatbot dialog to achieve better performance.

Keywords: human conversation, chatbot, Machine Learning, Natural Language Processing

1. Introduction

Chatbots are widely used in many common areas, software, and applications, from entertainment to business, including online shopping and social media. Therefore, chatbots can provide not only technical support in many different areas but also entertainment for users and customers. Chatbots can provide help for many users at the same time, which is more efficient and affordable compared to human customer service, which means less time and money cost to the companies. Aside from customer service support and assistance, chatbots can provide entertainment and companionship for the end user [1].

Chatbots make use of techniques and algorithms from two branches of AI: natural language processing and machine learning [2]. Firstly, a chatbot takes user input and processes it. Then it finally generates an output. Chatbots typically accept natural language text as input, and the final output should be the most pertinent result for the user's input sentence. Chatbots, therefore, represent an automated dialogue system that can support thousands of potential users at the same time [3].

With the help of recent advancements in artificial intelligence and approaches for natural language processing, chatbots can now mimic human speech more convincingly than ever before, are easier to create, and are more versatile in their use and upkeep. However, the human-chatbot interaction is not flawless. Contextual and emotional comprehension, as well as gender prejudice, are some areas that need development [4]. Chatbots are less capable than humans of understanding context and emotional and language clues, limiting their capacity to participate in more interesting and amicable interactions. At the same time, chatbots frequently behave stereotypically and take on traditionally feminine positions, doing them with traditionally female traits, demonstrating a gender bias in chatbot implementation and application. How to improve the implementation and evaluation of chatbots are still major research topics nowadays.

With the development of deep learning algorithms, the use of chatbots has grown [5]. The development of intelligent personal assistants is one of the newest and most fascinating applications (like Apple's Siri, Google's Google Assistant, Microsoft's Cortana, Amazon's Alexa, and IBM's

Watson) [6]. Chatbots, conversational agents, or personal assistants, are often integrated into smartphones, wearables, specialized home speakers and displays, and even automobiles, which can converse with the user through speech. For instance, when a user speaks a wake-up word or wake-up phrase, the chatbot wakes up and the intelligent personal assistant begins to listen [7]. The assistant can then comprehend orders and answer user queries by giving information, typically, using natural language understanding. In terms of technology, user interface, and functionality, all intelligent personal assistants have the same essential qualities [8]. The most advanced chatbots may also offer entertainment in addition to assistance with routine chores. Some chatbots have more developed personalities than others; the term "social chatbots" is used to describe these chatbots [9]. Despite recent advances in deep learning and natural language processing, there are still certain issues with chatbot designs. Due to a flawed dialog modeling strategy, the numerous language models offered for chatbot design still fall short of accurately simulating human discussions.

This paper uses the Cornell Movie Dialogs Corpus as the data set. This article examines how different model variables affect the performance of chatbots based on the attention-based neural machine translation model [10]. To evaluate the effectiveness of the chatbot, this article first created a question template with eight typical inquiries. At iteration 800, the chatbot only responds with the phrase "I don't know," indicating that it can still not comprehend and interpret the inquiries. The chatbot's vocabulary expands with each iteration. The solutions grow clearer, shorter, and more rational as n approaches 8000. The chatbot's responses to Q1, Q3, and Q4 are relevant to the queries. Numerous terms in the responses have nothing to do with the questions when the data volume of the train set is 20%. The train set can express grammatically accurate words and respond to some inquiries when its data volume is boosted by 50%. According to the overall experiment findings, raising the ITERATION and Movie line ratios can enhance the vocabulary and logic of chatbot discourse to provide higher performance.

2. Methods

2.1. Neural machine translation model

Attention-based neural machine translation model analyses two straightforward and useful groups of attention mechanisms: a comprehensive method that constantly takes into account all source words, and a local one that only takes a portion of the source words into account at once [11]. The model demonstrates the effectiveness of both techniques in both directions for the WMT translation processes between German and English. They also perform extensive analysis to assess their learning models, ability to deal with long sentences, choice of attentional structures, alignment effectiveness, and translation results. this paper selects this model as a baseline.

In this model, attentional decisions are made independently. Decisions on alignment in attentional NMTs should be made cooperatively, considering the alignment data from the past. A strategy that chains attention vectors to inputs used in subsequent phases, known as an input-feeding approach has been proposed. Instead of using context vectors in building hidden states, the model used in this paper adopts a broader strategy: the input-feeding approach, which is applicable to all recurrent stack designs, including those used by non-attentive models, and which gives the model the freedom to choose among many options without regard to attentional limitations.

2.2. Datasets

The WMT's 14 training data sets consisting of 4.5 million sentence pairs (116 million English words, 110 million German phrases) are used in the training process of the original models. They limit their vocabulary to the 50,000 most common words for both languages [12]. When training NMT systems, they exclude sentence pairings that are longer than 50 words, mixing mini-batches in the process. The four layers of the Stacking LSTM models each comprise 100 cells and 1000-dimensional embeddings [13]. The code was originally implemented in MATLAB. When run on a single Tesla K40 GPU device, a speed of 1,000 target words per second was reached. After collecting

both English-German and German-English results, In-depth analyses to comprehend the models' learning capacities for dealing with lengthy words, attentional architecture selection, and targeting quality were conducted. Eventually, the model builders concluded that in many circumstances, attentional-based NMT models outperform inattentive ones, such as translating names and handling complex and long sentences. This article selects Cornell Movie-Dialogs Corpus as a train set and retrains the attention-based neural machine translation model.

3. Experimental results and analysis

3.1. Dataset description

This paper uses the Cornell Movie Dialogs Corpus as the data set. To study the qualities of memorable movie phrases, a source for movie lines and a design for memorability are required. Hence, all of the lines from about 1000 films, which range in genre, age, and popularity, have been collected by the authors into a corpus. They then took the list of quotes for each movie from the relevant IMDbs Memorable Quotes page. The dataset focuses its comparisons on the memorable citations, which are presented as a single phrase rather than a block of many lines. For each (single sentence) memorable quote M, we identify a non-memorable quotation that is as identical to M as feasible in all respects except the selection of words. That means We want it to be the same length and voiced by the same character in the same movie. They select a quotation N from the same movie that contrasts with each M and is as close to the script M as feasible (either before or after). This corpus includes an extensive variety of fictional dialogues that were taken from uncut film screenplays. It contains 9035 characters from 617 films, making up 304713 total utterances.

3.2. Results and analysis

Based on the Attention-based neural machine translation model, this paper did further experiments to explore how different model variables affect the chatbot's performance.

Specifically, I changed the number of iterations of the training process and the size of the training data to find out how they affect the chatbot's performance respectively. I ran the code 7 times in total to test all pre-set conditions. This paper first built a simple question template containing eight common questions to test the chatbot's performance. The questions are shown in Table 1. Among them, Q1, Q2, Q3, Q4, Q6, Q7, and Q8 are open questions designed to test whether the chatbot can give logical and understandable answers. I also expect these questions could test its ability to understand the content of the questions. Q5 is a simple question and can be answered by yes or no. It is designed to test whether the chatbot can understand the questions, identify the type of the question, and make corresponding answers.

Table 1. Question template

Q1: What's your name?
Q2: How old are you?
Q3: What is your favorite color?
Q4: Where are you from?
Q5: Do you like listening to music?
Q6: What's your favorite movie?
Q7: How is the weather today?
Q8: What's your favorite city?

Apparently, when the number of iterations is 800, all answers the chatbot gives are "I don't know," meaning it still cannot understand and process the questions. When iteration changes from 800 to 2400, you can see that the chatbot's vocabulary is increasing, and the number of short answers is increasing. Although the logic is still confusing, it has more words for each question than the 800-iteration-training. When n_iteration changes from 2400 to 4000, We can see that the answers to some

questions are now more logical than before. Instead of simply starting all answers with “I don’t know”, now it has some more content in the answers. Especially in question two, the chatbot includes a number in its answer when the question is about its age. As n_iteration increases, so does the chatbot's vocabulary. When the n reaches 8000, the answers become more logical, understandable, and shorter. For Q1, Q3, and Q4, the answers given by the chatbot are highly related to the questions.

Apparently, when the number of iterations is 800, all answers the chatbot gives are "I don't know," meaning it still cannot understand and process the questions. When iteration changes from 800 to 2400, you can see that the chatbot's vocabulary is increasing, and the number of short answers is increasing. Although the logic is still confusing, it has more words for each question than the 800-iteration-training. When n_iteration changes from 2400 to 4000, We can see that the answers to some questions are now more logical than before. Instead of simply starting all answers with “I don’t know”, now it has some more content in the answers. Especially in question two, the chatbot includes a number in its answer when the question is about its age. As n_iteration increases, so does the chatbot's vocabulary. When the n reaches 8000, the answers become more logical, understandable, and shorter. For Q1, Q3, and Q4, the answers given by the chatbot are highly related to the questions.

Table 2. Change the Amount of ITERATION

	800	2400	4000	8000
Q1	I don’t know.	i don't know. i know. he's making him.	i don't know. men asleep. sorry. horse.	dr. david ravell.
Q2	I don’t know.	i don't know. dollars. you know my daughter.	thirty-five. things to do.	i don't know
Q3	I don’t know.	i don't know. i guess. you wanna do?	no. i want to see you. we?	Blue.
Q4	I don’t know.	i don't know. i did. sorry. sorry.	no. times. times. dollars. you?	southern southern
Q5	I don’t know.	no, no, no, no, no, no, no, no, no, no,	no. times. times. you -- times.	yes. i don't.
Q6	I don’t know.	i don't know. i don't know. sorry.	i don't know. one of them. course ya?	i don't know.
Q7	I don’t know.	i don't know. you know what happened.	i don't know. no idea. crazy.	i don't know.
Q8	I don’t know.	i don't know. i don't know. sorry.	i don't know. in a hurry.	blue.

Specifically, for question one, its answer is highly related to the question. Although the chatbot gives meaningless answers like "I don't know," the number of iterations is 8000, and its ability to identify the type of questions has a noticeable improvement. For example, it answers blue when asked about its favorite color instead of "I don't know" or any other irrelated words or phrases. For the question about where it came from, it answers a direction instead of a specific place which also has a significant improvement compared to the ones with fewer iterations.

Besides, as increase the number of interactions, it can also find out that the average loss continues to decrease. However, the model reached a threshold at 7300 iterations. It can see that the average loss finally stabilizes around 2.

Table 3. Change the Proportion of Movie_line

	20%	50%	80%
Q1	alexander	rufus.	jacob singer.
Q2	i'm eleven and a half	twenty.	twelve.
Q3	it's a secret jeff. a little thing.	blue.	don't worry.
Q4	southern california.	southern southern southern in san southern	in the park.
Q5	i don't know.	yes.	yes.
Q6	he's a great man, that's all.	don't know what you're talking about.	it's not a bad job in the morning.
Q7	fine.	it's a secret.	i don't know.
Q8	i don't know.	i don't know.	they'll be able to see you now.

When the data volume of the data set is 20%, there are many words in the answers, but they are not too related to the questions. When the data volume of the data set is increased to 50%, it can express grammatically correct sentences and answer some questions. When the data size of the data set increases to 80%, chatbots can answer most simple questions. The chatbot's understanding ability can be improved when the amount of data in the data set is increased. When the data volume of the data set increases to 100%, expect the chatbot to be able to provide high-quality responses.

Through the whole experiment results, increasing the number of ITERATION and increasing the proportion of data set can improve the vocabulary and logic of chatbot dialogue to achieve better performance.

4. Conclusion

Firstly, chatbots' background and recent development are discussed, and the Cornell Movie-Dialogs Cor-pus dataset is studied. In this paper, the effects of chatbot variables on its performance are experimentally set up, and eight rounds of dialogue templates are designed to observe the chatbot's answers to draw experimental conclusions. During the experiments, this paper mainly focuses on two variables that affect the final performance of the chatbot – the number of iterations and the training data size.

During the experiment, this paper first adjusts the number of iterations n_Iteration = 800. It then sets it to 2400, 4000, and 8000 to collect the dialogue results of the model. As N_Iteration = 800 increases, the chatbot's performance increases, and the chatbot's performance is better. Then this paper adjusted the data size of Movie_line and set three gradients, 20%, 50%, and 80%, and found that as the percentage of Movie_line increased, the chatbot's performance, including the ability to identify the type of questions, vocabulary, and logic also improved. The following work plan is to gradually subdivide the gradients, such as setting N_Iteration 400 to a gradient and Mov-ie_line 10% to a gradient, watch the chatbot performance change, protect the performance change trend, and find the best performance settings for the chatbot. Since the current code to train the chatbot runs relatively slower than expected, further algorithm optimization could also help this experiment be more efficient. Besides, a standard criterion to evaluate the quality of the chatbot's answers is also expected since many questions are open questions. A possible solution to this is that set up a scoring system from different aspects of a valid answer, including logic, understandability, fluency, pertinence length, word richness, and so on. With the possible adjustment of the question template, the related factors that affect performance could be more complex than the current work.

References

[1] Guendalina Caldarini, Sardar Jar, and Kenneth McGarry.: A Literature Survey Of Recent Advances in Chatbots, Information, 2022

[2] S. Nithuna and C. A. Laseena.: Review on Implementation Techniques of Chatbot, 2020 International Conference on Communication and Signal Processing (ICCSP), 2020, pp. 0157-0161

[3] Sharma, Mudita and Russell-Rose, Tony and Barakat, Lina and Matsuo, Akitaka.: Building a Legal Dialogue System: Development Process, Challenges, and Opportunities, arXiv, 2021, https://arxiv.org/abs/2109.00381

[4] Shum, Hy., He, Xd. & Li, D.: From Eliza to XiaoIce: challenges and opportunities with social chatbots., Frontiers Inf Technol Electronic Eng 19, 2018, pp10-26

[5] A. Shrestha and A. Mahmood.: Review of Deep Learning Algorithms and Architectures vol. 7, pp. 53040-53065, 2019, doi: 10.1109/ACCESS.2019.2912200.

[6] N. Goksel Canbek and M. E. Mutlu.: On the track of Artificial Intelligence: Learning with Intelligent Personal Assistants, HumanSciences, 2016, vol. 13, no. 1, pp. 592–601

[7] Russell Belk, Chatbots 2rd edition, Routledge, 2022

[8] Eleni Adamopoulou and Lefteris Moussiades.: Chatbots: History, technology, and applications, Machine Learning with Applications Vol 2, 2020, pp. 100006

[9] van Wezel, M.M.C., Croes, E.A.J., Antheunis, M.L.: "I'm Here for You": Can Social Chatbots Truly Support Their Users? A Literature Review, Springer International Publishing, 2021, pp. 96-113

[10] S. Sharma, M. Diwakar, P. Singh, A. Tripathi, C. Arya and S. Singh.: A Review of Neural Machine Translation based on Deep learning techniques, 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), 2021, pp. 1-5

[11] Luong, Minh-Thang and Pham, Hieu and Manning, Christopher D. .: Effective Approaches to Attention-based Neural Machine Translation, arXiv, 2015, https://arxiv.org/abs/1508.04025

[12] Garg, Ankush and Agarwal, Mayank.: Machine Translation: A Literature Review, arXiv, 2019, https://arxiv.org/abs/1901.01122

[13] Ghimire, Sujan and Deo, Ravinesh C. and Wang, Hua and Al-Musaylh, Mohanad S. and Casillas-Pérez, David and Salcedo-Sanz, Sancho.: Stacked LSTM Sequence-to-Sequence Autoencoder with Feature Selection for Daily Solar Radiation Prediction: A Review and New Modeling Results, Energies Vol 15, 2022, ISSN 1996-1073