Comparing methodologies to create an effective question-answer chatbot

By: Michael Wiciak

Contents

1 Introduction………………………………………………………………………………………………………………………………………….1

2 Research……………………………………………………………………………………………………………………………………………….3

3 Shortlisting Methodologies……………………………………………………………………………………………………………………5

4 Choosing a Model………………………………………………………………………………………………………………………………….7

5 Implementation…………………………………………………………………………………………………………………………………….9

6 Evaluation……………………………………………………………………………………………………………………………………………..9

[References 12](#_Toc57645970)

**1 Introduction**

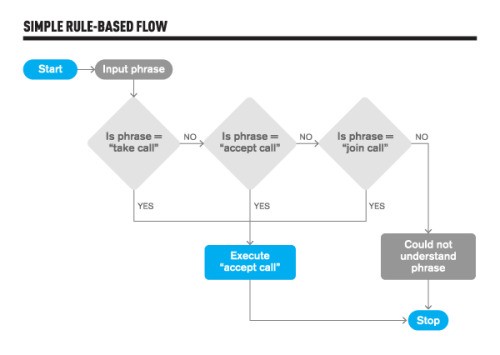
Chatbots are intelligent conversational systems that are able to process human language [25]. Chatbots such as ‘Vonage’ and ‘Domino's anywhere’ [1] are used widely in customer service, to improve customer satisfaction without the need of human-to-human interactions [2]. Chatbots can process the user input using the NLP[[1]](#footnote-1) tool, and then associate it with intent, to produce an output [3]. Rule-based chatbots are restricted to limited input options, giving rise to AI-based chatbots, which provide output considering variations of inputs; continuously learning and improving and creating its own rules. [4]

Figure : How rule-based chatbot work [4].

Yet, is it much harder to quantify how AI chatbots work. The methodologies they are based on, the functions they specialise in, for which scenario they are optimized for vary greatly. I am going to explore methodologies of chatbots, researching how they work, what they specialise in, and what their disadvantages are. My aim is to I find a methodology which seems to be most optimal for a question-answer chatbot based upon my research. I want my artefact to be able to effectively replicate a specific methodology, or at least do it in principle. It doesn’t matter what questions the chatbot can answer, as long as it follows the model it set out to try to replicate.

**2 Research**

Figure : Summarisation of different methodologies.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Methodology | How it works | Specialization | Real-life applications | Drawbacks |
| Watson Assistant [5] (WA) | Leverages AI capabilities to understand questions that user inputs in natural language. It uses machine learning models that are built from supplied data to deliver accurate and appropriate responses in real time. | Provides a smooth and factual conversation with a user. | Online banking and customer service chatbots. | The model requires a large quantity of training and validation data which needs to be 100% correct, or the model will learn incorrectly [25].  Might need to train for a longer time [6].  There are some instances where IBM Watson requires some complex coding for proper integration. [6] |
| RNN [7] | Input sentence is encoded into a sequence of vectors and then a subset of these vectors is chosen adaptively during the decoding translation. | Specialises in quick and accurate translations from one language to another one as the model is able to find a linguistically plausible soft alignment between the input sentence and the corresponding output sentence. | Machine translation,  robot control,  time series prediction,  speech recognition,  speech synthesis,  time series anomaly detection,  rhythm learning.[25] | Not able to handle unknown and rare words since only sentences consisting of known words are considered during translation [9]. NMT[[2]](#footnote-2) [8] tends to produce a short-sighted output. |
| Answer Selection Sentence [9] | ASS methodology applies distributional sentence models, in order to match questions with answers via considering the encoded semantics. | Specialises in making simple chatbots, which consider the ambiguity of each sentence and the double-meaning of each word. | Subtask of question- answering chatbots in customer service and remote therapy. | Inherit inability to learn words not included in the dataset or the training data.  [25] |
| Deep Reinforcement Learning for Dialogue Generation (DRLDG) [10] | Seq2Seq [11] model but with the  Maximum-likelihood estimation (MLE loss) [12] function. It is first trained with the usual MLE loss and then fine-tuned with policy gradients [13] to be optimized for specific conversational properties. | The model specializes on semantic coherence, in terms of mutual information between source and target. For questions, the response must make sense with respect to the query before being outputted. | Generating responses for conversational chatbots used in mental health therapies. | The problem lies with hardness in the evaluation of conversational agents since metrics such as perplexity and BLEU [14] do not always reward desirable conversational properties. [15] |
| SuperAgent Chatbot [16] | SuperAgent crawls the information of HTML and scrape PI, QA and CR[[3]](#footnote-3) data from the webpage, and then process the sub engines in parallel. If a high confidence answer is found from the first three engines, SuperAgent returns this answer as a response. Otherwise, the reply will be generated from the Chit Chat engine. | The accuracy of the FAQ search task for the proposed model surpasses the accuracy of the all previous models by a large margin. | This is a Chatbot that can answer anything, given enough data, and can become an independent AI chatbot, without human supervision. [25] | SuperAgent chatbot still cannot replace all the complex tasks which, at the moment, require human input.  [26] |
| CNN [17] | Convolutional neural network is a part of deep neural networks based on shared-weights architecture and translation invariance characteristics. | Specialises in images, videos processing and ranking candidate sentences. | Image processing, classification, segmentation and also for other auto correlated data [18] | CNN do not encode the position and orientation of the object into their predictions, making them forget any previous outputs it has given.  [19] |

Watson Assistant represents a standard machine learning cycle of training data, validation data [5]. This is common in most methodologies, like SuperAgent, RNN, DRLDG and ASS. It distinguished itself with the use of real-time data [25]. Other methodologies have additional enhancements: ASS adds semantics [9] into the data, while training DRLDG methodology [10] rewards the model prediction correct answers against the odds, RNN encodes the data into vectors [7] rather than tensors and decodes only a random batch rather than the whole training data. CNN differs as it is a type of a logistic regression [18] and specialises [17] on image prediction rather than text generation. Yet, SuperAgent creates 5 sub engines [26], each having a different task, and are used in parallel to find the best response. Comparatively, QA engine [5] is similar to standard Watson assistant methodology, CR engine [9] use similar semantic ideas like ASS and its Chit Chat engine uses NMT, which is based on Seq2Seq [11] model that is used extensively by RNN.



Figure : Shows how CNN works [17].

**3 Shortlisting Methodologies**

Both SuperAgent and Deep Reinforcement Learning for Dialogue Generation are models which require extensive and professional implementation [15][10], which makes them intractable to implement, whereas ASS methodology is only an improvement [9] to already existing model. By the process of elimination, the methodologies shortlisted are RNN, CNN and WA.

CNN Methodology [18] describes a feed-forward Neural Network that is mainly used to classify objects, due to its feature of breaking each component into subcomponents. Whereas in RNN Methodology [19] reconstructs components into a combination of different subcomponents. RNN [11] methodology is working on the principles of saving each output of a layer and then feed this back into the input while training it in a recurring loop. When used as a chatbot, CNN [17] methodology only considers the current input and it cannot remember the previous input, in contrast to a RNN methodology. The main different between CNN and RNN [19] is its use of connectivity patterns. CNN [25] uses connectivity patterns between its neurons, hence the neurons are arranged in such a way that enables CNN to respond to overlapping regions tilting the visual field. Whereas, RNN [7] uses time-series information, hence it is the most suitable model for systems that take conversation context in its considerations. On the other hand, Watson Assistant [5]operates by extracting data from a dataset, stemming the data [20], applying the data into a bag-of-words algorithm [21] which is often used by RNN [7], converting the data into a tensor[[4]](#footnote-4) and feed-forward the data. The model softmaxes [22] the data into probabilities, similar to CNN, and tries to predict responses according to inputs [17]. It learns by adjusting its weights and biases [5] after each cycle.

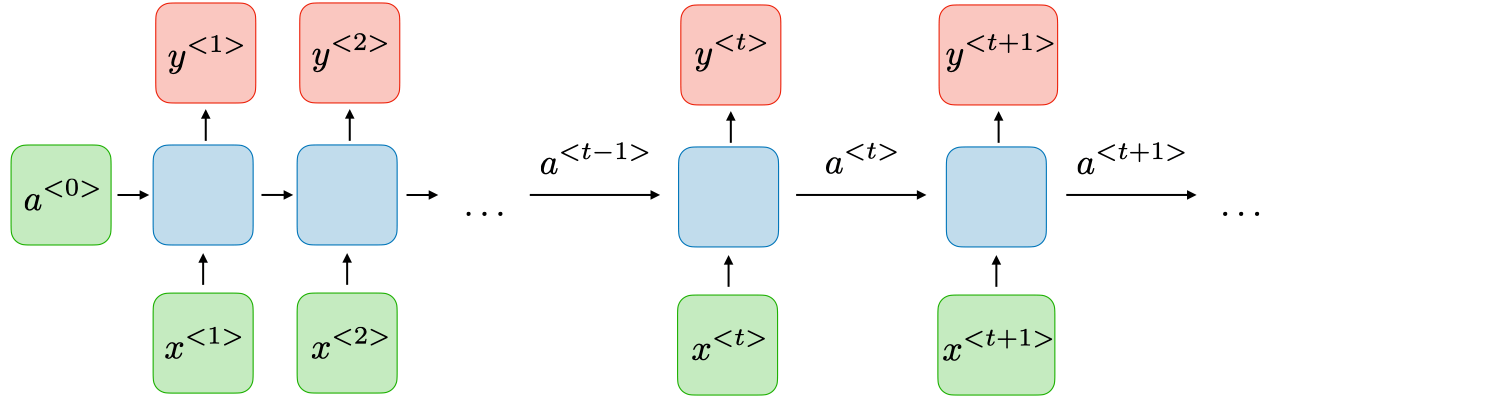


Figure : Shows how RNN works [7].

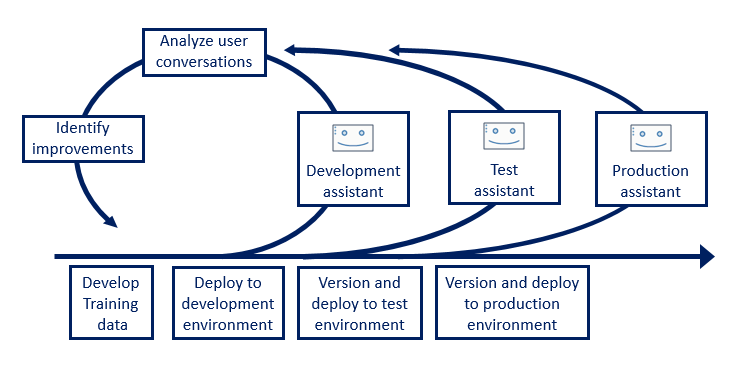


Figure : Shows the critical path of Watson Assistant methodology [5].

**4 Choosing a Model**

Although RNN methodology is effective in text/speech generation due to its inheritance of considering the context [7], the same result can be achieved by suppling a Watson Assistant a sparse/large range of data. IBM Watson methodology will notice the nuances in the training data and consider them when making a decision [5]. In such cases, RNN methodology will be much slower than an IBM Watson methodology.

Although CNN [18] specialises in finding factual information in text, WA could achieve the same result if all the information in the dataset is factually true. In such case, there is no need to distinguish whether the question/response is true.

CNN and RNN can train much faster than a Watson Assistant chatbot [6]. However, if time isn’t a constraint, Watson Assistant can outmatch/match both RNN and CNN. Therefore, the most optimal methodology is Watson Assistant.

WA methodology is widely used in industry today due to its efficiency [6]. The utilization of a Development Assistant, Test Assistant and Production Assistant, providing constant analysis of user conversations which helps to identify improvements in the model [5], makes it a reliable and constantly improving model, ideal for a question-answer chatbot. The use of training data makes it a relatively manageable model to implement.

**5 Implementation**

Below I implemented Watson Assistant methodology. The purpose of the chatbot is to recreate WA methodology, so content of the training dataset isn’t important. Dataset created I have created is based on Giant Pandas.

|  |  |  |
| --- | --- | --- |
| Function | Description | Code screenshot |
| Loading the a already trained model | If a model was already trained, it loads it from a file. If not, it indicates to train a new model. | C:\OneDrive\OneDrive - The Grammar School at Leeds\005784\Desktop\image.png |
| Loads libraries[[5]](#footnote-5)  [5] | Loads pre-defined libraries which ease with the implementation of the methodology. | C:\OneDrive\OneDrive - The Grammar School at Leeds\005784\Desktop\image.png |
| Formatting the data [5] | It collects training data generated and combines it into a numpy[[6]](#footnote-6) array[[7]](#footnote-7). | C:\OneDrive\OneDrive - The Grammar School at Leeds\image.png |
| Stemming the data [20] | Performs a pre-defined stemming algorithm, and assigns a 1 or 0 to data which recreates the responses associated with the tags. | C:\OneDrive\OneDrive - The Grammar School at Leeds\image.png |
| Softmax algorithm [22] | Applies a softmax algorithm which reduces data to probabilities. | C:\OneDrive\OneDrive - The Grammar School at Leeds\image.png |
| Bag of words algorithm [21] | Performs a bag of words algorithm, allowing the model to recognise similar words of the training data. | C:\OneDrive\OneDrive - The Grammar School at Leeds\thumbnail_image.png |

Figure : Explaining different parts of my program.

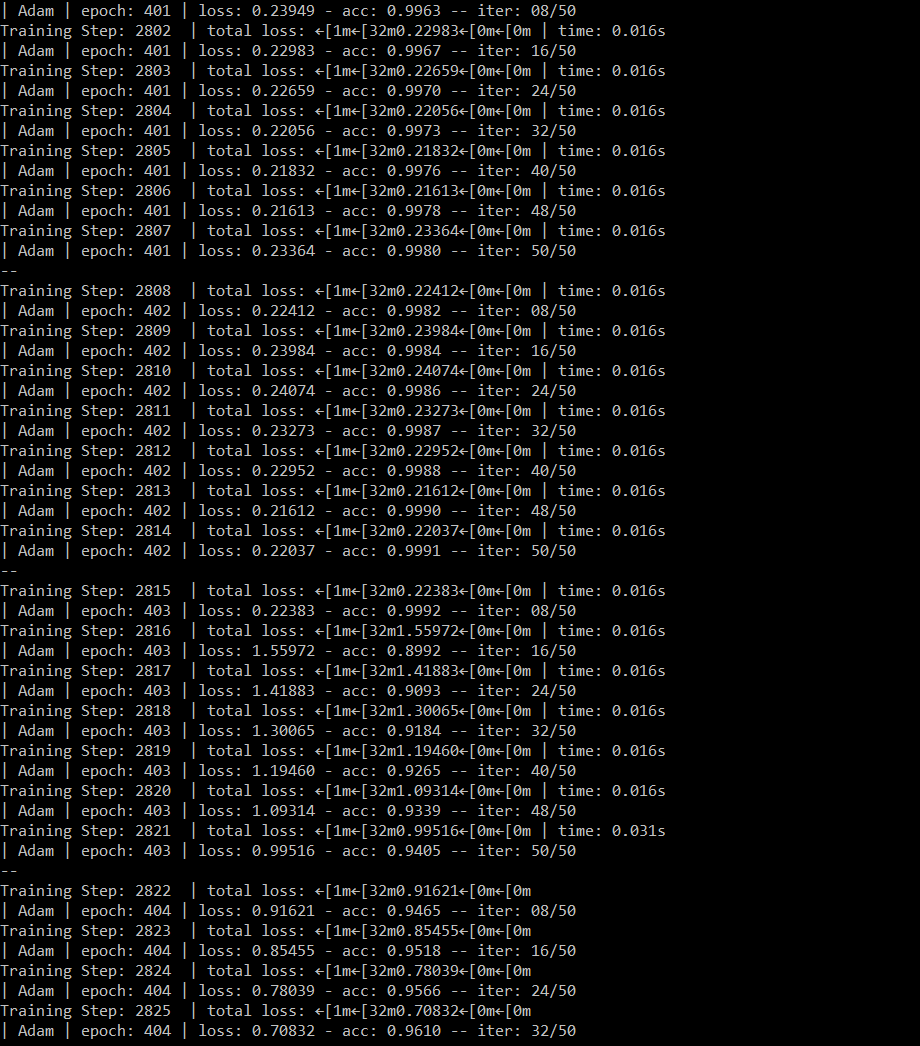
Figure : Functions and implementation of the chatbot.

Figure 7: Shows my chatbot mid-training

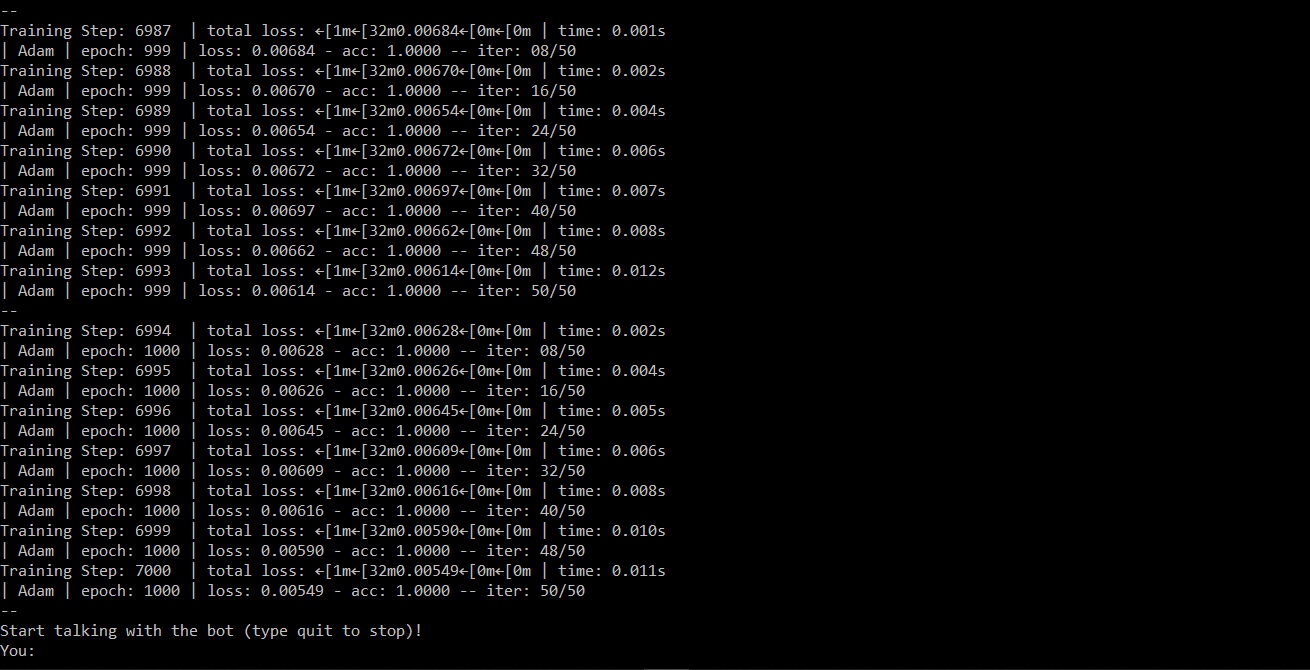
****

Figure 8: Chatbot at the end of training

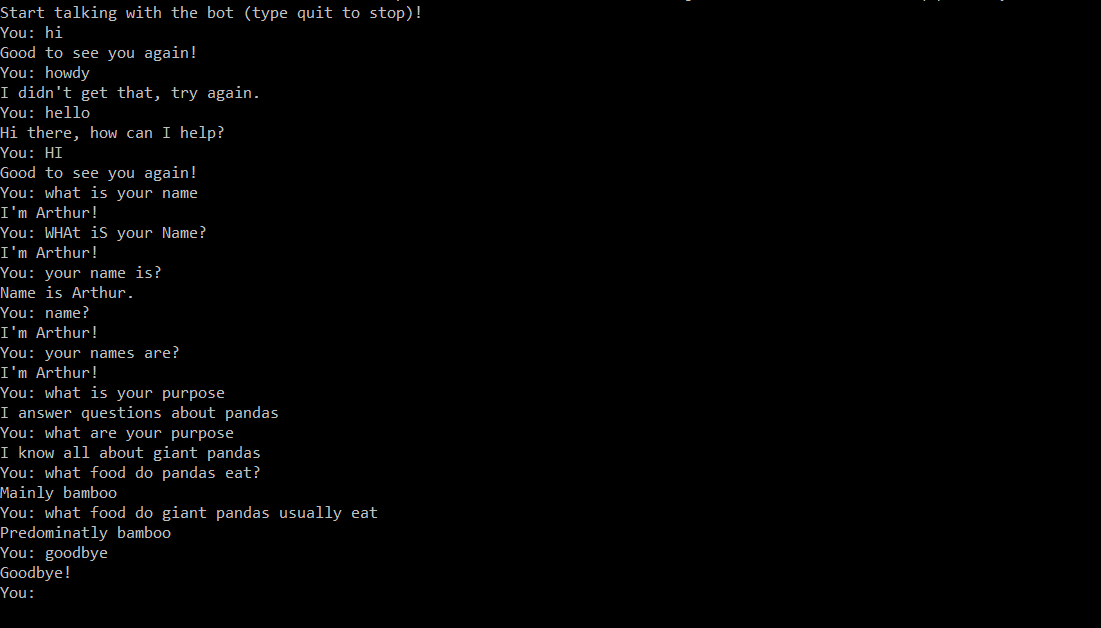
****

Figure 9: Shows my chatbot in operation.

**6 Evaluation**

Using research into machine learning algorithms, I have implemented the basics of WA methodology. My model takes training data, performs the standard algorithms WA would. However, my model is still a simplification. WA outranks other chatbots in accuracy due to effective training cycle it performs by training through a shear amount of data [25]. It takes large datasets such as Squad2.0 [23] and puts them into the model. Yet, the dataset used only contains a small amount of questions on a singular topic, which is opposite of what Watson Assistant would be supplied. My model does not match the accuracy of WA, removing some of its advantages over other methodologies. WA dynamically learns, meaning it learns even when the training of it is complete, allowing it to be flexible to new questions and environments. Although my model keeps track of every conversation it had and loads from file, it does not perform dynamic machine learning like Watson Assistant does.

To conclude Watson Assistant is the most effective chatbot methodology if given enough data. From personal experience implementation of all its functions is difficult yet, it is much easier to implement compared to methodologies such as CNN, which require an extensive deep neural networks [19].

# References

|  |  |
| --- | --- |
| [1] | M. Grech, “5 Great Examples of Customer Service Chatbots in Action,” Get Voip, 09 04 2019. [Online]. Available: https://getvoip.com/blog/2019/04/09/customer-service-chatbots/. [Accessed 10 11 2020]. |
| [2] | D. Shewan, “10 of the Most Innovative Chatbots on the Web,” WordStream, 26 02 2020. [Online]. Available: wordstream.com/blog/ws/2017/10/04/chatbots#:~:text=Chatbots%20–%20also%20known%20as%20“conversational,based%20applications%20or%20standalone%20apps.. [Accessed 10 11 2020]. |
| [3] | A. Bolen, “What are chatbots?,” SAS Institute, 2020. [Online]. Available: https://www.sas.com/en\_gb/insights/articles/analytics/what-are-chatbots.html#:~:text=Using%20chatbots%2C%20computers%20can%20understand,Language%20Analytics%20Strategist%20at%20SAS.. [Accessed 10 11 2020]. |
| [4] | K. Shridhar, Artist, *Rule based bots vs AI bots.* [Art]. Medium, 2017. |
| [5] | IBM, “Development process,” 24 04 2020. [Online]. Available: https://cloud.ibm.com/docs/assistant?topic=assistant-dev-process. [Accessed 10 11 2020]. |
| [6] | Trust Radius Reviews, “IBM Watson Assistant Reviews,” Trust Radius, 18 01 2020. [Online]. Available: https://www.trustradius.com/products/ibm-watson-assistant/reviews?qs=pros-and-cons. [Accessed 10 11 2020]. |
| [7] | A. Amidi, “Recurrent Neural Networks cheatsheet,” 12 02 2020. [Online]. Available: https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks. [Accessed 10 11 2020]. |
| [8] | J. Brownlee, “A Gentle Introduction to Neural Machine Translation,” 29 12 2017. [Online]. Available: https://machinelearningmastery.com/introduction-neural-machine-translation/. [Accessed 10 11 2020]. |
| [9] | K. M. Hermann, “Deep Learning for Answer Sentence Selection,” Department of Computer Science University of Oxford,, Oxford, 2014. |
| [10] | J. Li, “Deep Reinforcement Learning for Dialogue Generation,” Department for Computational Linguistics, Vancouver, 2016. |
| [11] | F. Chollet, “A ten-minute introduction to sequence-to-sequence learning in Keras,” 29 09 2017. [Online]. Available: https://blog.keras.io/a-ten-minute-introduction-to-sequence-to-sequence-learning-in-keras.html. [Accessed 10 11 2020]. |
| [12] | J. Brownlee, “Loss and Loss Functions for Training Deep Learning Neural Networks,” 28 01 2019. [Online]. Available: https://machinelearningmastery.com/loss-and-loss-functions-for-training-deep-learning-neural-networks/. [Accessed 10 11 2020]. |
| [13] | J. Peters, “Policy gradient methods,” 05 11 2010. [Online]. Available: http://www.scholarpedia.org/article/Policy\_gradient\_methods#:~:text=Policy%20gradient%20methods%20are%20a,cumulative%20reward)%20by%20gradient%20descent.. [Accessed 10 11 2020]. |
| [14] | J. Brownlee, “A Gentle Introduction to Calculating the BLEU Score for Text in Python,” 27 11 2017. [Online]. Available: https://machinelearningmastery.com/calculate-bleu-score-for-text-python/. [Accessed 10 11 2020]. |
| [15] | I. V. Serban, “Off-policy REINFORCE with Learned Reward Function,” in *A Deep Reinforcement Learning Chatbot*, Canada, School of Computer Science, McGill University, 2017, pp. 21-22. |
| [16] | L. Cui, “SuperAgent: A Customer Service Chatbot for E-commerce Websites,” 04 08 2017. [Online]. Available: https://www.aclweb.org/anthology/P17-4017.pdf. [Accessed 15 11 2020]. |
| [17] | G. Zaid, “Methodology for Efficient CNN Architectures in Profiling Attacks – Extended Version,” 19 11 2019. [Online]. Available: https://eprint.iacr.org/2019/803.pdf. [Accessed 10 11 2020]. |
| [18] | J. Brownlee, “When to Use MLP, CNN and RNN Neurnal Networks,” Towards Data Science, 23 07 2018. [Online]. Available: https://machinelearningmastery.com/when-to-use-mlp-cnn-and-rnn-neural-networks/. [Accessed 10 11 2020]. |
| [19] | F. L, “Machine Learning - RNN vs CNN at a high level - Data Science Stack Exchange,” 2016. [Online]. Available: https://datascience.stackexchange.com/questions/11619/rnn-vs-cnn-at-a-high-level. [Accessed 10 11 2020]. |
| [20] | M. Porter, “The Porter Stemming Algorithm,” Tartarus , 01 2006. [Online]. Available: https://tartarus.org/martin/PorterStemmer/. [Accessed 10 11 2020]. |
| [21] | J. Brownlee, “A Gentle Introduction to the Bag-of-Words Model,” Machine Learning Mastery, 09 10 2017. [Online]. Available: https://machinelearningmastery.com/gentle-introduction-bag-words-model/. [Accessed 10 11 2020]. |
| [22] | Uniqtech, “Understand the Softmax Function in Minutes,” Data Science Bootcamp, 30 01 2018. [Online]. Available: https://medium.com/data-science-bootcamp/understand-the-softmax-function-in-minutes-f3a59641e86d#:~:text=Softmax%20is%20exponential%20and%20enlarges,encoded%20in%20one%20hot%20encoding).. [Accessed 10 11 2020]. |
| [23] | M. Li, “Question Answering on SQuAD2.0,” 06 06 2020. [Online]. Available: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/default/15816213.pdf. [Accessed 08 11 2020]. |
| [24] | R. Chandra, “What are the disadvantages of a recurrent neural network?,” Quora, 27 10 2018. [Online]. Available: https://www.quora.com/What-are-the-disadvantages-of-a-recurrent-neural-network. [Accessed 10 11 2020]. |
| [25] | R. Kumar, “A Review on Chatbot Design and Implementation Techniques,” *International Research Journal for Engineering and Technology (IRJET),* vol. III, no. 11, pp. 2791-2811, 2020. |
| [26] | L. Cui, “SuperAgent: A Customer Service Chatbot for E-commerce Website,” in *55th Annual Meeting of the Association for Computational Linguistics-System Demonstractions*, Vancouver, 2017. |

1. 1 Natural language processing is a branch of artificial intelligence that helps computers understand, interpret and manipulate human language. [↑](#footnote-ref-1)
2. Neural Machine Translation [8] [↑](#footnote-ref-2)
3. Fast Questions and Answers, Product Information, Customer Reviews [↑](#footnote-ref-3)
4. 4 Algebraic object that describes a relationship between sets of algebraic objects related to a vector space. [↑](#footnote-ref-4)
5. Libraries are a collection of precompiled routines that a program can use. [↑](#footnote-ref-5)
6. Library often used in machine learning [↑](#footnote-ref-6)
7. Data structure which consists of elements, identified by an index. [↑](#footnote-ref-7)