ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 1, 2024

INTELLIGENT CHATBOT USING DEEP LEARNING BY PYTHON

G.Uma Maheswari , Professor, Department Of Data Science, SICET, Hyderabad

P.Sai Srikar Reddy,J.Sai Kiran,T.Ajay,N.Shiva Koti Reddy,Pavan Kumar Reddy

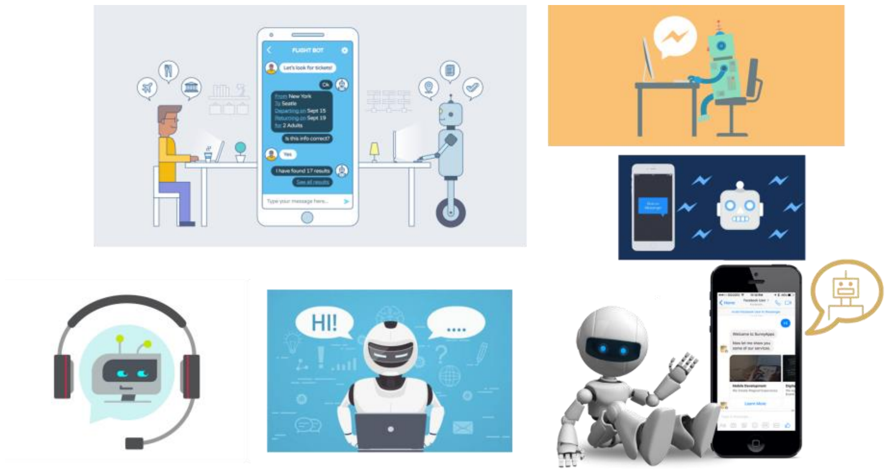
UG Student, Department Of Data Science, SICET, Hyderabad

Abstract

Conversation generation or intelligent conversational agent development using artificial intell igence or machine learning is an interesting problem in natural language processing. In many R&D projects, they use artificial intelligence, machine learning algorithms and natural langua ge techniques to create interactive/conversational systems. Their research and development c ontinues and testing continues. Negotiation brokers are often used by businesses, government agencies, and nonprofit organizations. These are often used by financial institutions such as b anks and credit card companies, as well as businesses such as online retailers and startups. Th ese virtual employees are used by many businesses, from small startups to large corporations. There are many policy-based and interface-based chatbot developments on the market. But they lack the flexibility and practicality to en gage in a real conversation. Popular personal assistants include Amazon's Alexa, Microsoft's Cortana, and Google's Google Assistant. These workers are limited to work, are returning wo rkers, and are not designed to engage in conversations that simulate human relationships. Ma ny of the existing chatbots were developed using rulebased methods, simple machine learning algorithms, or accessbased techniques, but these techniques fail to produce beautiful results. In this project, I developed an intelligent interactive agent using cuttingedge techniques sugge sted in a recent research paper. To build intelligent chatbots, I used Google's Neural Machine Translation (NMT) model, which is based on the sequencetosequence (Seq2Seq) model with an encoderdecoder model. This encoderdecoder uses a recurrent neural network with bidirecti onal LSTM (longterm memory) units. For efficiency, I use the neural listening mechanism an d beam searching during training.

Introduction

A chat agent or chatbot is a program that generates responses based on instructions to simulat e human interaction in text or speech. These applications are designed to simulate human inte raction. Chatbots are mostly used in business and commercial organizations, including gover nment, nonprofit organizations, and private organizations. Their responsibilities range from customer service, product advice, product inquiries to personal assistants. Many interactive agents are designed using rulebased tools, mining technologies, or simple machine learning algorithms. In accessbased technology, the chat agent scans the input message for keywords and retrieves relevant responses based on the query. They are based on similar content, and the text is extracted from internal or external sources, including the World Wide Web or corporate repositories. Some other advanced chatbots are developed using natural language processing (NLP) and machine learning algorithms. Additionally, many business chat engines exist to create chatbots based on customer profile input.



Recently, there has been great interest in the use and dissemination of the new generation communication process. Many large technology companies use virtual assistants or chat agents to meet customer needs. These include Google's Google Assistant, Microsoft's Cortana, and Amazon's Alexa. Although mostly Q&A, their adoption by major companies has increased customer satisfaction and appears to promise a facetoface meeting to oversee research and development. Related WorkChat agent has seen a lot of development and testing lately. In addition to traditional chatbot development techniques that use rulebased techniques or simple machine learning, many advanced chatbots use natural language processing (NLP) and deep learning techniques such as deep neural network (DNN) and deep learning (DRL). The sequenctosequence (Seq2Seq) model, based on the standard encoderdecoder architecture, is one such model that is very popular in networking, modeling, and machine translation. Seq2Seq uses neural networks (RNN), a popular deep neural network architecture designed specifically for langua

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 1, 2024

ge processing tasks. In the sequencetosequence (Seq2Seq) model, manytomany RNN architec ture is used for the decoder. In this encoderdecoder architecture, the input sequence is fed to t he encoder as a vector representation of the text. The encoder then creates some intermediate representation of the message or thought vector. Therefore, the thought vector produced by th e encoder is fed as input to the decoder. Finally, the decoder creates thought vectors and trans forms the sequences one by one, producing many outputs from the decoder in the form of tar get sequences. Although ordinary RNN is used by default in Seq2Seq and can effectively sol ve many NLP problems, due to the complexity of the speech structure, physical units often fa il, especially when longterm data needs to remember data, which occurs frequently. It will be larger than larger data and will have insufficient data trust of RNN network. That's why resea rchers use the evolution of neural networks to solve these problems.

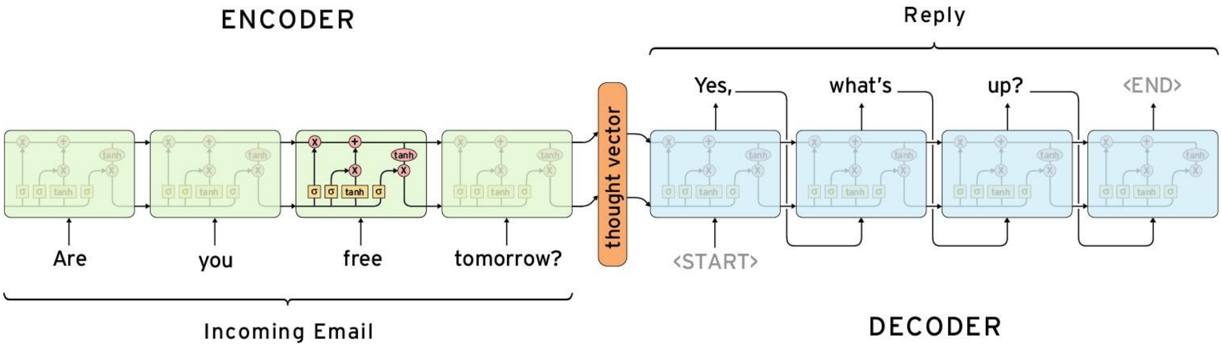


Figure 2.1: Sequence to Sequence Model

Shortterm memory (LSTM) is a special type of neural network cell that has experience prove n useful for modeling language. In addition to input and output gates, LSTM also has a memo ry gate. This will help remember important information and content and clear out the entire s ystem; this is best in language modeling as dependency in array is not uncommon. Additional ly, bidirectional LSTM units may be more efficient than unidirectional units. So we are follo wing industry standard practice. In the neural listening mechanism, each hidden target is com pared with the base hidden state, a maintenance vector is created by calculating the scores, an d the color vector is stored in memory to select another candidate. Additionally, other method s, such as beam searching, can also help improve the decisionmaking process by selecting the

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 1, 2024

best candidates. Seq2Seq has also been used in other NLP tasks, including machine translation, text recognition, response, and image recognition.

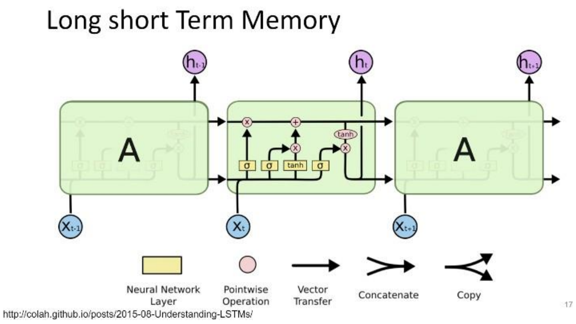


Figure 2.2: Ref 7

2.2 Google's Neural Machine Translation (GNMT)

Google's Neural Machine Translation (GNMT) model is a model for neural machine translation between other languages and English. GNMT has also been used experimentally for intergenerational communication. It is based on the popular Seq2Seq model of session generation. Additionally, the GNMT module incorporates many technologies required for the development of intelligent chatbots. GNMT models include linktosegment models of encoderdecoder architectures designed using either unidirectional or bidirectional LSTM units. They also have neural listening mechanisms, beam search, and word generation options using Google's subword module. They can also choose to tune hyperparameters for better model training. ReinRequired Learning (DRL) was used to create long conversation chatbots.

They do a great job with the Q&A and the interview format is great too. But it is not possible to test the real interaction of people and it does not have a simple function. Some chatbots that use machine learning algorithms often follow simple algorithms. They lack the knowledge and skills needed to be successful in speaking clearly. These are also black boxes, and business customers have poor visibility into their internal devices. Therefore, the results they produce may not meet the customer's expectations

4. Deep neural network for chatbots

4.1 Recurrent neural network

Recurrent neural network is a special deep neural network architecture used mainly for natura l language processing (NLP). In normal deep neural networks, the memory or part of the data is not calculated. However, in recurrent neural networks, sequence information is stored in memory and used for further processing, making RNN suitable for continuous data or time recording by decision. > A recurrent neural network (RNN) consists of an input layer, multiple hidden layers, and an output layer. In the input layer, the input is given as a vector representation. The input vector is then divided by some weight and some additional bias. The output of the input layer is then passed to the next hidden layer, where each primitive layer contains multiple RNNs. After the output is received from the input process, the units in the process are divided into outputs from the input process according to their weights and deviations. Then, in each hidden class, some global functions (sigmoid, tangent) are used to generate the output of the hidden process. The output of each hidden unit is then passed to the hidden process. Similar to the previous secret room, some weight, bias and activation has been applied to the ideas of the current secret room. This process is revealed by all subsequent hidden processes. Finally, the output of the last hidden layer is sent to the output layer, which uses some functions (like Softmax) to produce the final output. For RNN, the output vector of the final output is fed into the feedback process instead of the input vector. Therefore the data array is stored and used in memory.

However, vanilla RNN stores data sequentially. For large files with long segments, this can result in poor data on the network. It may cause reduced network performance due to data overload. In many cases, the data set may not be relevant to many NLP tasks, including speech generation, leading to poor modelling. This problem is solved by a special type of RNN unit long term memory (LSTM). Neural network unit that solves the data bottleneck in long arrays. In addition to input and output gates, LSTM also has a memory gate. This helps the system remember longer without overloading the network by providing less information.

The challenge of creating a chatbot or electronic conversation engine is to create interactive communication. Since the model used in this experiment is for machine translation, the speech generation process, in which the history of the previous speech is not included, is treated as a translation problem. Therefore, the effectiveness of the model in long-term communication will be limited. Another challenge is finding the right hyperparameters to optimize the translation of the chatbot or conversation generation process. GNMT is a selfexplanatory model with bidirectional LSTM units, neural listening mechanisms, and channel s earch technology. Many of these features improve machine translation speech generation. Careful bidirectional LSTM units tend to produce better output. The Seq2Seq module also has some advantages of GNMT. However, it would be a better choice to create chatbot algorithms from scratch by developing RNN, bidirectional LSTM and neural listening strategies because GNMT is generally used for machine translation. However, this requires a lot of trial and error to achieve an effective chatbot mode and is therefore more appropriate than a research question. But most of the output is repetitive and generic. Additionally, due to the lack of realife data, chatbot performance is lower than the best human interaction in practice. Additionally, many messages were deleted due to length or inconsistency. In addition, if the number of training instructions is less than necessary and the comparability of test and development data will negatively affect the operation of the model. Additionally, due to limited data, longterm training may not be suitable for structured discussions. DiscussionCreating interactive agents using Neural Machine Translation (NMT) is widely used. Some of the other methods are to just use the combined pattern. Many people also use their own sequences for Sequence modes. But they work poorly due to lack of complexity. However, the communication problem can be better solved by spending more time and effort to create more interactive communication. Therefore, ray tracing may be a better job with array-to-array based encoderdecoder architectures that rely on bidirectional LSTM units and neural listening mechanisms. Improve optimization if better data is available. Before using the Cornell movie dataset, I tested other datasets and pretrained them on the GNMT module. However, they were later excluded from training due to lack of good data. Additionally, the information available here are video captions that rarely include human interaction. More realistic, reallife interactive data can simulate users' needs and personalities. For future training, personal conversation history can be combined to give the chatbot personality. However, some of the responses were repetitive and lacked value. This can be reduced by adding different and healthy ingredients. Additionally, adding more length to older posts can help improve replies and make them more relevant. Due to the length limit option, messages over 100 are discarded but the full text is retrieved later, resulting in most items being lost. This may lead to repetition in the training process. Additionally, repeated words of the same character are removed, except for the speaker's last words, which are repeated.This further reduces the file size. Therefore, more data with longer dimensions can help create smarter chatbots. Future WorkChatbots built using Google Neural Machine Translatio

ISSN2454-9940
www.ijasem.org
Vol 18, Issue 1, 2024

n Models (GNMT) can be further enhanced with powerful, reaworld chatbots that better simulate affected human interaction. Additionally, the hyperparameters of the GNMT model can be further optimized and finetuned to improve performance. Depending on the way to expand the task, deep learning (RL) can be used, which can improve performance, as seen in Dufarsky's paper. Reinforcement learning algorithms can be used after initial training with Google Neural Machine Translation

Conclusion

The educational effect of the Cornell film subtitle structure needs to be further improved, and more knowledge and emphasis should be given to the teaching parameters. Adding higher quality data will improve performance. Additionally, the training model needs to be trained with other hyperparameters and different data for further testing. This is an attempt to use deep neural networks for speech generation to create intelligent chatbots. Deep learning for the interactive generation.

REFERENCES

Abelson, H., Sussman, G. J., & Sussman, J. (1996). ***Structure and interpretation of computer programs*** (2nd ed.). The MIT Press. https://mitpress.mit.edu/sites/default/files/sicp/

Allaire, J., Xie, Y., Dervieux, C., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2024). *rmarkdown*___: Dynamic documents for R___. https://github.com/rstudio/rmarkdown

Anderson, J. R. (1990). ***The adaptive character of thought***. Lawrence Erlbaum.

Anscombe, F. J. (1973). Graphs in statistical analysis. ***The American Statistician***, ***27***(1), 17–21. https://doi.org/10.2307/2682899

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. ***Science***, ***211***(4489), 1390–1396. https://doi.org/10.1126/science.7466396

Bache, S. M., & Wickham, H. (2022). *magrittr*___: A forward-pipe operator for R___. https://magrittr.tidyverse.org

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. ***Acta Psychologica***, ***44***(3), 211–233. https://doi.org/10.1016/0001-6918(80)90046-3

Bar-Hillel, M., & Falk, R. (1982). Some teasers concerning conditional probabilities. ***Cognition***, ***11***(2), 109–122. https://doi.org/10.1016/0010-0277(82)90021-X

Bateman, S., Mandryk, R. L., Gutwin, C., Genest, A., McDine, D., & Brooks, C. (2010). Useful junk? The effects of visual embellishment on comprehension and memorability of charts. ***Proceedings of the SIGCHI Conference on Human Factors in Computing Systems***, 2573–2582. https://doi.org/10.1145/1753326.1753716

Baumer, B. S., Kaplan, D. T., & Horton, N. J. (2021). *Modern Data Science with R* (2nd ed.). Chapman; Hall/CRC. https://mdsr-book.github.io/mdsr2e/

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. ***Psychological Methods***, ***2***(2), 131–160. https://doi.org/10.1037/1082-989X.2.2.131

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? ***Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency***, 610–623. https://doi.org/10.1145/3442188.3445922

Bergstrom, C. T., & West, J. D. (2021). ***Calling bullshit: The art of skepticism in a data-driven world***. Random House Trade Paperbacks. https://www.callingbullshit.org/

Bertin, J. (2011). *Semiology of graphics: Diagrams, networks, maps* (Vol. 1). ESRI Press.

Billings, Z. (2021). *bardr:* ***Complete works of William Shakespeare in tidy format***. https://CRAN.R-project.org/package=bardr

Binder, K., Krauss, S., & Wiesner, P. (2020). A new visualization for probabilistic situations containing two binary events: The frequency net. ***Frontiers in Psychology***, ***11***, 750. https://doi.org/10.3389/fpsyg.2020.00750

Box, G. E. (1979). Robustness in the strategy of scientific model building. In ***Robustness in statistics*** (pp. 201–236). Elsevier. https://doi.org/10.1016/B978-0-12-438150-6.50018-2

Box, G. E. P. (1976). Science and statistics. ***Journal of the American Statistical Association***, ***71***(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Cairo, A. (2012). *The functional art: An introduction to information graphics and visualization*. New Riders.

Cairo, A. (2016). *The truthful art: Data, charts, and maps for communication*. New Riders.

Chang, W. (2012). ***R graphics cookbook: Practical recipes for visualizing data*** (2nd ed.). O'Reilly Media. https://r-graphics.org/

Cleveland, W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. ***Science***, ***229***(4716), 828–833. https://doi.org/10.1126/science.229.4716.828

Craik, K. J. W. (1943). ***The nature of explanation***. Cambridge University Press.

Crameri, F., Shephard, G. E., & Heron, P. J. (2020). The misuse of colour in science communication. ***Nature Communications***, ***11***(1), 1–10. https://doi.org/10.1038/s41467-020-19160-7

Davenport, T. H., & Patil, D. J. (2012). Data scientist: The sexiest job of the 21st century. ***Harvard Business Review***, ***90***(5), 70–76. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

De Cruz, H., Neth, H., & Schlimm, D. (2010). The cognitive basis of arithmetic. In B. Löwe & T. Müller (Eds.), *PhiMSAMP. Philosophy of mathematics: Sociological aspects and mathematical practice* (pp. 59–106). College Publications. http://www.lib.uni-bonn.de/PhiMSAMP/Data/Book/PhiMSAMP-bk_DeCruzNethSchlimm.pdf

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R., et al. (2017). Curriculum guidelines for undergraduate programs in data science. ***Annual Review of Statistics and Its Application***, ***4***, 15–30. https://doi.org/10.1146/annurev-statistics-060116-053930