Singlish Chatbot

NTU-AI6127: Project Final Report

Brandon Chua Shao Jie
School of Computer Science and Engineering
Nanyang Technological University
`BRAN0026@e.ntu.edu.sg`

Chen Yongquan
School of Computer Science and Engineering
Nanyang Technological University
`S200043@e.ntu.edu.sg`

Clarence Teoh Ken Guan
School of Computer Science and Engineering
Nanyang Technological University
`S200035@e.ntu.edu.sg`

Fong Lin Qiang
School of Computer Science and Engineering
Nanyang Technological University
`FONG0127@e.ntu.edu.sg`

Lam Wai Hong
School of Computer Science and Engineering
Nanyang Technological University
`S190101@e.ntu.edu.sg`

Abstract

Singlish is a unique form of language commonly used by average Singapore nationals. Its mix of different languages - English and the several ethnic languages - makes it a challenge for language model to understand and interpret. This project aims to develop a chatbot that understand Singlish and enable meaningful interactions with the older generation in Singapore who may not be well-versed with emerging technologies. The Singlish chatbot will be built using a transform-based attention model.

1 Introduction

The free communication between humans and machines is the ultimate objective of natural language processing (NLP) [1]. Multiple studies on NLP have agreed that the progress of information exchange between these two parties have been contributed heavily by chatbots [2]. According to [3], chatbots are defined as automated systems that aid users by responding to their queries. Several industrial sectors including advisory, service, and commercial have already employed the use of chatbots [4]. As proposed by [5], chatbots are required to have the following requirements such as having conversational capabilities to engage with the users rather than simply giving a yes or no response, responding with information that is semantically correct and meaningful for ease of understanding [6], and be trained on a specific domain of data [7] so that it can comprehend the specific terms and information. To satisfy the above requirements, the modelling of chatbots to consistently provide accurate responses to match the context of the users queries remains a challenge.

In our work, we aim to develop a chatbot with the ability to understand and converse in Singlish. A unique form of language commonly spoken by the average Singapore national, Singlish is a mix of different languages which borrows its vocabulary from English and the several ethnic languages found in Singapore. The purpose of developing these chatbots is to enable Singaporeans, especially those who are not fluent in English such as the older generation, to have meaningful interactions with these AI systems in their own language of familiarity. However, while our literature search may not have been exhaustive, we have not found any published study (in scientific journals) that has trained a chatbot to learn Singlish. To address this novel challenge, we have deployed a transformer-based attention model to learn this unique language.

An encoder-decoder model using Long Short Term Memory (LSTM) serves as the baseline. Such a model, also known as a sequence-to-sequence (seq2seq) model, is able to convert sequences from one domain to another. Hence, it is not only useful for machine translation tasks, but also text generation tasks. It is in this respect, that such a seq2seq model is designated as a base model on which other models will be compared.

2 Related Work

2.1 Singlish Datasets & Work

Singlish as a low resource English creole language is not entirely novel to the linguistic community. *Cambria et al.* in 2015 proposed using a combination of support vector machine and a custom algorithm incorporating n-gram Singlish sentic patterns and multilingual polarity sentic patterns for polarity sentiment analysis for Singlish text. They collected tweets on topics relevant to Singapore from Singaporean users on Twitter [8], constructing a Singlish-English polarity lexicon using Singlish terms extracted from these tweets. *Cambria et al.* later on added Singlish support into the SenticNet concept-level resource for sentiment analysis [9]. *Wang et al.* also created their own treebank for investigating dependency parsing of Singlish [10]. They crawled for Singlish posts from the *SG Talk Forum*[1], manually tagging and annotating the sentences under the Universal Dependencies scheme. This dataset was later used for later work done on studying transfer learning for Singlish universal dependency parsing and part-of-speech tagging [11], and also parsing Singlish with multingual cased BERT [12]. However, this dataset is not publically available. *Chen et al.* also collated a dataset of SMS messages from Singaporeans and students from the National University of Singapore, splitted into a Singaporean English corpus and a Mandarin Chinese corpus [13]. The Singaporean English corpus includes Singlish sentences but they are not specifically ordered according to coherent dialogues and only includes the sentences from the sender. However, it is still possible to artificially construct possible dialogue sessions by matching sentences within the corpus amongst each other.

2.2 Dialogue Generation

Approaches towards building chatbots or dialogue generation models can be mainly categorized into rule-based and data-driven approaches [14]. ELIZA [15] and PARRY [16] are instances of rule-based chatbots that uses pattern and transform rules for matching responses to prompt utterances. They are fast and easy to build compared to data-driven models but rules can become highly complex for performant chatbots. Data-driven models can be further information retrieval based and machine learning based models. Information retrieval based chatbots relies on information retrieval techniques such as using word-level vector space models and cosine distance for finding the most probably question-response pair [17]. Term Frequency-Inverse Document Frequency (TF-IDF) retrieval models have also been proposed [18, 19]. Meanwhile, machine learning approaches branch into neural networks based and reinforcement learning based. For the latter, it has been proposed that dialogue sessions be modeled as Markov Decision Processes (MDPs) [20] or Partially Observed Markov Decision Processes (POMDPs) [21, 22, 23, 24].

For this project, we will be focusing on neural network based approaches for building the Singlish chatbot. While the 2014 sequence-to-sequence (Seq2Seq) encoder-decoder architecture broke new grounds in word-level language modelling [25], modelling conversational dialogue presents greater challenges. Seq2Seq originally recurrent neural networks like Long Short-Term Memory (LSTM) [26] or gated recurrent units (GRU) [27]. With the introduction of the attention mechanism [28, 29], replacing the recurrent units with Transformers [30] pushed the boundaries even further and are used in the current state-of-the-art models like GPT [31, 32, 33], BERT [34] and BART [35]. While word-level language modelling for tasks like translation only require the model to predict the next word in context, sentence level dialogue modelling requires the model to predict subsequent words with relevance to the entire prompting sentence or historical dialogue. While a closed-domain expert system can sufficiently manage dialogue with a complex enough ruleset, open-domain dialogue modelling suffer from content or style inconsistencies [36, 37, 38], lack of long-term contextual information [39] and blandness [40, 41, 42]. Such problems are noted by Vinyals and Le when

---

[1] http://sgtalk.com/

they trained their Seq2Seq architecture using LSTM for dialogue modelling [43]. Transformers being better than LSTMs and GRUs at modelling long-term dependency information generally lend themselves to modelling conversations, thus DialoGPT [44] seeked to extend GPT-2 [32] for the dialogue modelling task. *Wu et al.* used a pretrained BERT as the encoder and GPT as the decoder [45]. Alibaba used an attentive Seq2Seq model for both generating responses and reranking its own responses and responses generated from a information retrieval model [46]. Meena [47] investigated using the Evolved Transformer [48] as part of a Seq2Seq model on multi-turn open-domain dialogue modelling. *Mehri et al.* proposes masked context retrieval and inconsistency utterance identification as pretraining objectives for dialog context representation learning [49]. Most recently, DialogBERT [50] proposed a hierachical Transformer architecture and two novel training objectives, them being masked utterance regression and distributed utterance order ranking as additional loss terms, allowing the model to better capture dialogue coherence. They compared their model using BART, DialoGPT as baselines. They also used a RNN model and a Transformer model, both using *Mehri et al.* pre-training methods, in their baseline models. They evaluated the models using perplexity, BLEU-4 and NIST, noting that perplexity correlated with human judgment scores significantly [47].

3 Approach

3.1 Data Preparation

The data used was a collection of SMS messages. More details pertaining to the data can be found in Section 4.1.

We first normalise the data into a tabular dataframe. We clean the text component of the data by removing non-ascii characters, chinese characters and multiple spaces and newlines. Text with less than 3 words and originating from non-Singapore users are filtered out. This ensures the sufficient text length to capture patterns and limit the language to Singlish by Singaporeans respectively. Additionally, we employ stemming and lemmatisation techniques.

In preparing the sentences for training, we give every sentence a preceding number of sentences that served as context. Thus, the amount of context provided for each response is a parameter that could determine the performance of the model. In this demonstration, we have maintained a number of 7 pieces of context per response. The data preparation phase is slightly different for the seq2seq baseline model, where the sentences are arranged as a sequence of questions and responses, akin to a dialogue format. Every response in turn acts as a question, and vice versa. Due to the large memory needed for the model, we use only the first 32,000 pairs. Of these, 20% are used as a validation set.

3.2 Model

The DialoGPT model [44] is used for this project. Its architecture is based on the GPT-2 model [32], which is an upscaled GPT model - a large transformer-based language model. GPT-2 proved that transformer-based models trained on very large datasets are able to capture long-term dependencies in textual data. Hence, the DialoGPT extends GPT-2 to address the challenges of conversational neural response generation.

DialoGPT is a 12-layer transformer with layer normalization and byte pair encodings for the tokenizer. Multi-turn dialogue sessions are concatenated into long text with source at the beginning portion of the text and target at the end. It uses masked self-attention that does not peek into the tokens to its right.

4 Experiments

4.1 Data

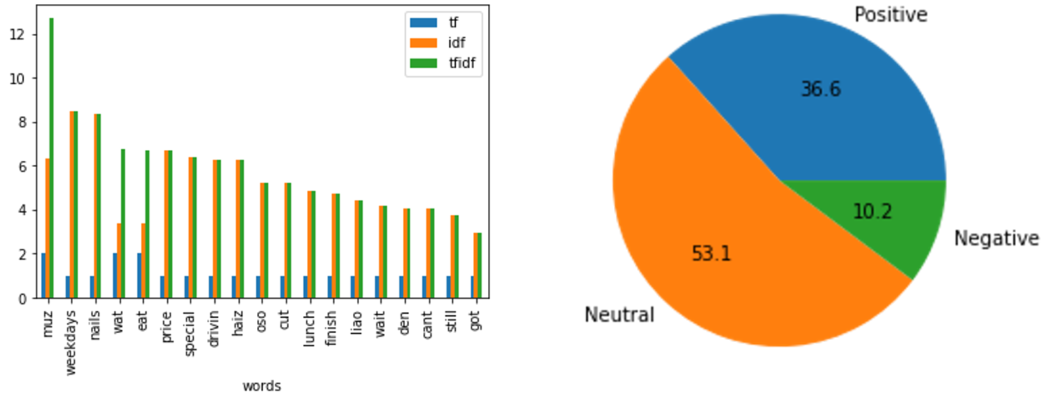
The dataset used in our work is the National University of Singapore SMS Corpus [13]. This dataset consists of 67093 SMS (Short Message Service) messages, with 35598 messages in Singapore English, and the remaining 31465 messages in Mandarin Chinese. These messages were collected from subjects participating in a research study, most of whom were Singaporeans attending the university. In our work, we will only be using the Singapore English (also known as Singlish) dataset to build our chatbot.

Before cleanup	Den only weekdays got special price... Haiz... Cant eat liao... Cut nails oso muz wait until i finish drivin wat, lunch still muz eat wat
After cleanup	Den only weekdays got special price Haiz Cant eat liao Cut nails oso muz wait until i finish drivin wat lunch still muz eat wat
Lowercase and removal of stop words and punctuation	den weekdays got special price haiz cant eat liao cut nails oso muz wait finish drivin wat lunch still muz eat wat

Table 1: Sample text data from Corpus

Our aim is to build a chatbot. The output, which is a proper conversational response, will be generated by the model based on the input sentence given by the user. Before using the dataset to train, pre-processing is carried out on the data. Firstly, simple clean up is performed such as removing non-ASCII characters, websites, newline, one-word characters, digits and extra white spaces. For the purpose of data analysis, the data is then converted to lowercase, with stop words and punctuation removed. Example:

TF-IDF is then carried out along with sentiment analysis.



(a) TF-IDF plot of a SMS text

(b) Pie chart of sentiment analysis of all SMS texts

Figure 1

4.2 Experimental Results using DialoGPT

Table 1 shows the hyper-parameters tested, where the first row are the hyper-parameters used for the base model. The best perplexity is given using the hyper-parameters at the third row from the bottom of the table. Unlisted values are identical to those of the base model.

Epochs	LR	Weight decay	Grad. Acc. Steps	Grad. clipping	Optimizer	Val set perp.
5	0.00005	0.0		1.0	AdamW	2.3525
				0.0		2.3525
				2.0		2.3526
			10			12.7012
			100			41.6984
		0.1				2.0938
		0.01				2.3219
	0.0001					491.9594
	0.001					1.3728
	0.01					536.3946
	0.1					984.6895
		0.1			SGD	292.5464
		0.1			SGD & 0.98 mom	158.8381
		0.1			Adamax	21.9244
	0.001	0.1				1.1459
3						5.5897
10						1.2026

Table 2: Hyper-parameters and Experimental Results

An example output of conversation with best perplexity score (1.1459):

» User: Oi, later meet bukit batok 6 Beng Bot: will be late think i ll eat dinner in school
» User: then liddat meet 8 at the arcade? Beng Bot: no rush going to be l
» User: wat u mean no rush, u the one who wanted to go arcade one leh Beng Bot: u there taking photo
» User: ya i taking photo your taiji meh Beng Bot: heee thanxoxo
» User: thank ur head ar, so how meet 8 at arcade isit Beng Bot: u went offline ar
» User: i as online as it gets liao bro Beng Bot: s ex la he said jus plan stay at e beach nothing else planned yet then u think need bring clothes chang
» User: of course lar, u want smelly smelly take mrt meh Beng Bot: haha okok gd gd luck 4r i got home le i use bd
» User:i thot u still at school?? Beng Bot: ya lor haha best is ben 2nd
» User: ben gonna smack u also if u dont tell us what time meet at the arcade Beng Bot: eh i ask my dad fetching me back to lot one then i take bus heh
» User: so isit got time to dinner or not Beng Bot: if u going ikea str wher

Table 3

Training with different pre-processed data set:

Epochs	lr	Weight decay	Gradient Acc. Steps	Gradient clipping	Optimizer
5	0.001	0.1	1	1.0	AdamW

When processing the data set, various methods were experimented on. One of the methods is to only use SMS texts of a certain word count. Another method is to use stemming and/or lemmatising. The results have shown that whether using stemming or lemmatising or both does not make much of a difference in perplexity. On the other hand, using a larger number of word count does improve perplexity. One thing to note is that the number of training examples are lesser when the word count required is higher. Therefore, the improvement in perplexity is likely due to the model overfitting due to smaller dataset. The following table shows the results:

For word_count > 3 (23482 training examples):

PorterStemmer	LancasterStemmer	WordNetLemmatizer	Validation set perplexity
Yes	No	Yes	1.1455
No	Yes	Yes	1.1512
Yes	No	No	1.1434
No	Yes	No	1.1470
No	No	Yes	1.1459

Table 4

For word_count > 6 (17192 training examples):

PorterStemmer	LancasterStemmer	WordNetLemmatizer	Validation set perplexity
Yes	No	Yes	1.0974
No	Yes	Yes	1.0985
Yes	No	No	1.0994
No	Yes	No	1.0982
No	No	Yes	1.0949

Table 5

For word_count > 10 (10926 training examples):

PorterStemmer	LancasterStemmer	WordNetLemmatizer	Validation set perplexity
Yes	No	Yes	1.0699
No	Yes	Yes	1.0702
Yes	No	No	1.0699
No	Yes	No	1.0702
No	No	Yes	1.0703

Table 6

Example output of conversation when word_count > 3:

» User: Oi, later meet bukit batok 6 Beng Bot: will be late think i ll eat dinner in school
» User: then liddat meet 8 at the arcade? Beng Bot: no rush going to be l
» User: wat u mean no rush, u the one who wanted to go arcade one leh Beng Bot: u there taking photo
» User: ya i taking photo your taiji meh Beng Bot: heee thanxoxo
» User: thank ur head ar, so how meet 8 at arcade isit Beng Bot: u went offline ar
» User: i as online as it gets liao bro Beng Bot: s ex la he said jus plan stay at e beach nothing else planned yet then u think need bring clothes chang

Table 7

Example output of conversation when word_count > 10:

» User: Oi, later meet bukit batok 6 Beng Bot: i is sleeping haha i just woke up p
» User: slp wut slp, wat time alr, dinner leh Beng Bot: dunno leh i random suggest suddenly all super on haha since u alsonot very interested in that movie and wq wanna watch also but gvpopcorn not nice
» User: ur stupid idea to watch one now u pangseh, bo steady leh Beng Bot: lol not bad arh but is hai ren one u won t like i tink lol okie night too
» User: ur stupid idea to watch one now u pangseh, bo steady leh Beng Bot: lol not bad arh but is hai ren one u won t like i tink lol okie night too
» User: u watch already U Still JIO US WATCH? ABIT guo fen ar dog Beng Bot: hey happy cny i don t know that s why was asking u all in email
» User: bro u high or what? dont come tok stupid with limpei leh Beng Bot: hee yes dearie dearest laogong sleep tighttight hold me tight mug mug
» User: muack ur head lar bye Beng Bot: hey i think got link just follow instruction only u need for

Table 8

During pre-processing, we can also choose to set the number of previous SMSes responses as replies for context. After experimenting with different values, we found out that using 7 previous replies gives the best perplexity score.

n number of previous replies	Validation set perplexity
7 (Base)	1.1459
3	4.9641
5	1.4063
9	1.6771

Comparing the DialoGPT model with the seq2seq model, we note that it has outperformed the seq2seq model. Keeping the data cleaning parameters similar, the perplexity of the seq2seq model is higher than the DialoGPT model, indicating a lower performance. The figures are summarized in the table below.

	DialoGPT	Seq2seq
Minimum Number of Word Count in SMS	3	3
Number of Previous Replies	7	1
Number of Epochs	5	5
Learning Rate	0.001	0.001
Optimizer	AdamW	Adamax
Perplexity Score	1.15	2.75

5 Analysis

It is important to choose the right hyper-parameters during training. Using a learning rate that is too small causes the model to learn too slowly while using a large gradient causes unstable training. Using more epochs improves the model but also increases training time. By using suitable parameters, we can train using less time and still achieve better results as shown in Table 2, where training for 5 epochs is able to give a better perplexity score than training for 10 epochs.

As mentioned in section 4.2, there is an improvement in perplexity when we use data sets that have word counts of less than 6 and 10 filtered out (Table 5 & 6). It has been pointed out that due to a smaller data set, there is a likely case of over-fitting. In general, using a large enough data set is vital to training a good model. Under the same section, using 7 previous replies as context for training gave the best perplexity. The reason 7 replies works better may be because it is the average length of SMS conversations. Using too little replies does not give the model enough context to work with. On the other hand, using too many decreases the accuracy of the context since the previous replies may belong to another conversation.

6 Conclusion

This project used the DialoGPT model for the construction of a Singlish Chatbot. From the example conversations shown in this report, it is clear that there are more work to be done. However, for start of conversation for word count more than 3 and more than 10, the first few replies by the chatbot seemed to make consist of common sense. Only at about 3rd or 4th turns, does the chatbot diverge from the conversation topic and stop speaking sense. This can be a good starting point for future works on building a Singlish-based chatbot.

The SMS Corpus used for this project encompasses the use of Singlish. However, it also included the use of short hand like LOL for "laugh out loud" and BTW for "by the way", which may not be relevant in the Singlish context where the focus is on Singlish words like "lah", "lor", "leh" etc. These short hands are a major factor when training the model. Another issue with managing such data is the standardisation of spelling. Using off-the-shelf lemmatizers like that in the Python NLTK package will not cover many of words unique to the Singlish language. Future work on creating a Singlish chatbot should focus on building up a proper conversational dataset that omits short hands. A customised lemmatizer should also be built from ground up so that the data can be processed properly in context.

References

[1] Sihyung Kim, Oh-Woog Kwon, and Harksoo Kim. Knowledge-Grounded Chatbot Based on Dual Wasserstein Generative Adversarial Networks with Effective Attention Mechanisms. *Applied sciences*, 10(9):3335, 2020.

[2] Jenhui Chen, Obinna Agbodike, and Lei Wang. Memory-Based Deep Neural Attention (mDNA) for Cognitive Multi-Turn Response Retrieval in Task-Oriented Chatbots. *Applied sciences*, 10(17):5819, 2020.

[3] Mohammad Nuruzzaman and Omar K. Hussain. IntelliBot: A Dialogue-based chatbot for the insurance industry. *Knowledge-based systems*, 196:105810, 2020.

[4] Petter Bae Brandtzaeg and Asbjørn Følstad. Why people use chatbots. In Ioannis Kompatsiaris, Jonathan Cave, Anna Satsiou, Georg Carle, Antonella Passani, Efstratios Kontopoulos, Sotiris Diplaris, and Donald McMillan, editors, *Internet Science*, pages 377–392, Cham, 2017. Springer International Publishing.

[5] Vladimir Vlasov, Johannes E. M. Mosig, and Alan Nichol. Dialogue Transformers. *CoRR*, abs/1910.00486, 2019.

[6] Adam Di Prospero, Nojan Norouzi, Marios Fokaefs, and Marin Litoiu. Chatbots as Assistants: An Architectural Framework. CASCON ’17, page 7686, USA, 2017. IBM Corp.

[7] Florian Daniel, Maristella Matera, Vittorio Zaccaria, and Alessandro Dell’Orto. Toward Truly Personal Chatbots: On the Development of Custom Conversational Assistants. In *2018 IEEE/ACM 1st International Workshop on Software Engineering for Cognitive Services (SE4COG)*, pages 31–36, 2018.

[8] Siaw Ling Lo, Erik Cambria, Raymond Chiong, and David Cornforth. A multilingual semi-supervised approach in deriving Singlish sentic patterns for polarity detection. *Knowledge-Based Systems*, 105:236–247, 2016.

[9] Danyuan Ho, Diyana Hamzah, Soujanya Poria, and Erik Cambria. Singlish SenticNet: A Concept-Based Sentiment Resource for Singapore English. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1285–1291, 2018.

[10] Hongmin Wang, Yue Zhang, GuangYong Leonard Chan, Jie Yang, and Hai Leong Chieu. Universal Dependencies Parsing for Colloquial Singaporean English. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1732–1744, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[11] Hongmin Wang, Jie Yang, and Yue Zhang. From Genesis to Creole Language: Transfer Learning for Singlish Universal Dependencies Parsing and POS Tagging. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(1), May 2019.

[12] Ethan C. Chau, Lucy H. Lin, and Noah A. Smith. Parsing with Multilingual BERT, a Small Corpus, and a Small Treebank, 2020.

[13] Tao Chen and Min-Yen Kan. Creating a Live, Public Short Message Service Corpus: The NUS SMS Corpus. *CoRR*, abs/1112.2468, 2011.

[14] Maali Mnasri. Recent advances in conversational NLP: Towards the standardization of Chatbot building, 2019.

[15] Joseph Weizenbaum. ELIZA - A Computer Program for the Study of Natural Language Communication between Man and Machine. *Commun. ACM*, 9(1):3645, January 1966.

[16] Kenneth Mark Colby. *Artificial Paranoia: A Computer Simulation of Paranoid Processes*. Elsevier Science Inc., USA, 1975.

[17] Rafael E. Banchs and Haizhou Li. IRIS: A Chat-oriented Dialogue System based on the Vector Space Model. In *Proceedings of the ACL 2012 System Demonstrations*, pages 37–42, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[18] Sudeep Gandhe and David Traum. Surface Text based Dialogue Models for Virtual Humans. In *Proceedings of the SIGDIAL 2013 Conference*, pages 251–260, Metz, France, August 2013. Association for Computational Linguistics.

[19] Franck Charras, Guillaume Dubuisson Duplessis, Vincent Letard, Anne-Laure Ligozat, and Sophie Rosset. Comparing System-response Retrieval Models for Open-domain and Casual Conversational Agent. In *Second Workshop on Chatbots and Conversational Agent Technologies*, Los Angeles, United States, 2016.

[20] E. Levin, R. Pieraccini, and W. Eckert. Using Markov decision process for learning dialogue strategies. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, volume 1, pages 201–204 vol.1, 1998.

[21] Nicholas Roy, Joelle Pineau, and Sebastian Thrun. Spoken Dialogue Management Using Probabilistic Reasoning. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 93–100, Hong Kong, October 2000. Association for Computational Linguistics.

[22] Jason D. Williams and Steve Young. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422, 2007.

[23] Steve Young, Milica Gai, Blaise Thomson, and Jason D. Williams. POMDP-Based Statistical Spoken Dialog Systems: A Review. *Proceedings of the IEEE*, 101(5):1160–1179, 2013.

[24] M. Gai, C. Breslin, M. Henderson, D. Kim, M. Szummer, B. Thomson, P. Tsiakoulis, and S. Young. On-line policy optimisation of Bayesian spoken dialogue systems via human interaction. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8367–8371, 2013.

[25] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to Sequence Learning with Neural Networks, 2014.

[26] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Comput.*, 9(8):17351780, November 1997.

[27] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, 2014.

[28] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate, 2016.

[29] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation, 2015.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, 2017.

[31] A. Radford and Karthik Narasimhan. Improving Language Understanding by Generative Pre-Training. 2018.

[32] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. 2019.

[33] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, 2020.

[34] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019.

[35] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, 2019.

[36] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. A Persona-Based Neural Conversation Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany, August 2016. Association for Computational Linguistics.

[37] Yizhe Zhang, Xiang Gao, Sungjin Lee, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. Consistent Dialogue Generation with Self-supervised Feature Learning, 2020.

[38] Xiang Gao, Yizhe Zhang, Sungjin Lee, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. Structuring Latent Spaces for Stylized Response Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1814–1823, Hong Kong, China, November 2019. Association for Computational Linguistics.

[39] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues, 2016.

[40] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A Diversity-Promoting Objective Function for Neural Conversation Models, 2016.

[41] Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. Generating Informative and Diverse Conversational Responses via Adversarial Information Maximization, 2018.

[42] Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. Conversing by Reading: Contentful Neural Conversation with On-demand Machine Reading. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5427–5436, Florence, Italy, July 2019. Association for Computational Linguistics.

[43] Oriol Vinyals and Quoc Le. A Neural Conversational Model, 2015.

[44] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation, 2020.

[45] Qingyang Wu, Lei Li, Hao Zhou, Ying Zeng, and Zhou Yu. Importance-Aware Learning for Neural Headline Editing, 2019.

[46] Minghui Qiu, Feng-Lin Li, Siyu Wang, Xing Gao, Yan Chen, Weipeng Zhao, Haiqing Chen, Jun Huang, and Wei Chu. AliMe Chat: A Sequence to Sequence and Rerank based Chatbot Engine. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 498–503, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[47] Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. Towards a Human-like Open-Domain Chatbot, 2020.

[48] David R. So, Chen Liang, and Quoc V. Le. The Evolved Transformer, 2019.

[49] Shikib Mehri, Evgeniia Razumovskaia, Tiancheng Zhao, and Maxine Eskenazi. Pretraining Methods for Dialog Context Representation Learning, 2019.

[50] Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. DialogBERT: Discourse-Aware Response Generation via Learning to Recover and Rank Utterances, 2020.