Exploring the effect of dataset on chatbot performance

Anonymous ACL submission

Abstract

This paper explored the effect of dataset on chat bot performance. The chat bot are trained on Cornell Movie Corpus, Daily Dialog Corpus, and the mix of Cornell Movie Corpus and Daily Dialog Corpus. In order to improve the performance of the chat bot, we trained a Dialog Act Classifier to label Cornell Movie Corpus. Then add Dialog Act as a feature to train the Chat bot. We evaluated the chat bot in (1) grammaticality and (2) naturalness (3) interestingness for a sample of 100 for the three different models.

1 Introduction

The use of conversational agents or a ChatBot, which are computer programs using natural language interact with human users, have become a trend in industry given advantages they bring about to our daily life. The main job they provide is automatic customer services, which reduces a large amount of human labors. Despite of huge attentions paid on the development of a ChatBot, there still some limitations that need to be improved. That is, most of the ChatBot models are designed to respond to questions and generate an appropriate answers in a restricted domain. Thus, the respond generated from the ChatBot is unnatural or not human-like. This is because training datasets for the Chatbot model is insufficient. As an attempt to improve this limitation, we try expanding an existing dataset for the Chatbot model. We implement a pytorch (**?**) ChatBot tutorial to Cornell Movie Corpus (**?**) and Daily Dialogue dataset(**?**) individually. Also, we combine the two datasets and apply it to the ChatBot model.

2 Related work

Rule-based or template-based methods (Williams and Zweig, 2016), (Wen et al., 2016) and dialogue state tracking are typically adopted close-domain systems (Henderson, 2015)(Wang and Lemon, 2013)(Wen et al., 2016). In contrast, data-driven techniques such as Seq2Seq generation are used for open-domain chatbots. In general, QA knowledge base or conversational corpus is used to train the Seq2Seq based generation chatbots to generate a response for each input(Wu et al., 2016). Several previous works reveal that RNN based Seq2Seq models are suitable for this work (Cho et al., 2014) (Sutskever et al., 2014) (Ritter et al., 2011)(Shang et al., 2015) (Sordoni et al., 2015) (Serban et al., 2016). (Sutskever et al., 2014) proposed a basic seq2seq model and other works such as (Bahdanau et al., 2014)(Sordoni et al., 2015) (Song et al., 2016) (Quarteroni and Manandhar, 2007) (Qiu et al., 2017) (Ghose and Barua, 2013) enhanced model with attention, context information and diversified answers. Although lots of work have done, the output of seq2seq generation models tend to be unrelated to input and senseless.

inputencwu2016sequential

3 Dataset

3.1 Cornell Movie Corpus

We use Cornell Movie Corpus, which contains a large collection of fictional conversations extracted from raw movie scripts. To be more specific, it is composed of 220, 579 dialogues between 10,292 pairs of characters in 617 movies, which involve the 9,035 characters. In total, there are 304, 713 utterances in the corpus. Features included in movie metadata are genres, release year, IMDB (Internet Movie Database) rating, and number of IMDB votes. Features of characters meta-

Dataset	number of conversation	dialogue act
Cornell Movie corpus	220,579	null
Daily Dialog	13,118	manually labeled
Cornell + Daily	233,697	classifier labeled

Table 1: Information of the dataset

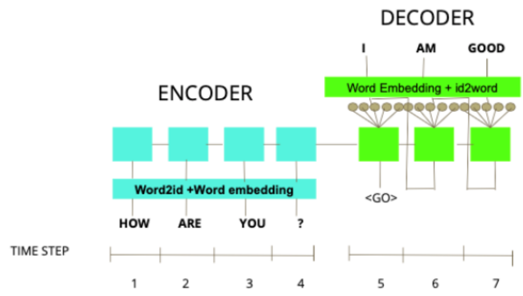
data include gender (for 3,774 characters) and position on movie credits (for 3,321 characters).

3.2 Daily Dialog

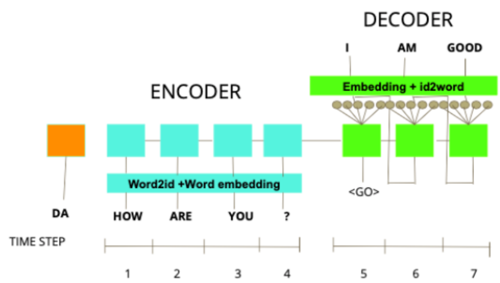
We also use Daily Dialogue dataset, which contains 13,118 multi-turn dialogues. This dataset is constructed by crawling the raw data from various websites where English learners practice English dialogue in daily life. Therefore, this dataset is written by human, which makes it more formal compared to other datasets, such as Twitter Dialog Corpus and Chinese Weibo dataset. Also, Daily Dialogue dataset includes conversations regarding with a certain topic, such as shopping and trips. For example, it includes a conversation between a customer looking for a particular product and a staff at a shop helping the customer. Also, it contains a conversation between two students talking about vacation trips. Moreover, dialogues in this dataset ends after more speaker turns compared to other datasets. That is, the dialogues in Daily Dialogue include in average about 8 turns, but about three topics in other datasets. When it comes to the average, average speaker turns per dialogue is 7.9, average tokens per dialogue is 114.7, and average tokens per utterance is 14.6. Also, the Daily Dialogue dataset is manually labeled to reflect intention of communication and human emotions. For intention of communication, which our project is focused on, each utterance in the dataset is labeled with one of four dialogue act classes, that is, Inform, when a speaker is providing information, Questions when a speaker is seeking for information, Directives when a speaker requests, instructs, suggest and accepts or rejects offer, and Commissives when a speaker accepts or rejects a request/suggestion/offer.

3.3 Mixed dataset

We first implement a chatbot model to Cornell Movie Corpus and Daily Dialogue dataset individually. In other words, we have a Cornell Movie Corpus, which is a dialogue dataset without a Dialogue Act (DA) label, and Daily Dialogue dataset, which already is already labeled with DA. After deleting DA from Daily Dialogue dataset, we combine Cornell Movie Corpus and Daily Dialogue as one dataset.



(a) Sequence to Sequence model



(b) Sequence to Sequence model and dialog act

Figure 1: Chat bot model

4 Sequence to Sequence Dialogue Agent

4.1 Data preparation

Handle loading and preprocessing of Cornell Movie-Dialogs Corpus dataset and daily dialogue dataset.

4.2 Implement a sequence-to-sequence model with Luong attention mechanism(s)

Luong attention used top hidden layer states in both of encoder and decoder. In Luong attention they get the decoder hidden state at time t. Then calculate attention scores and from that get

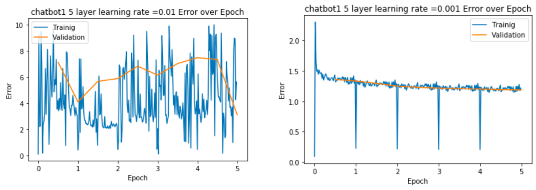


Figure 2: Learning rate 0.01 and 0.001 on chat bot 1

the context vector which will be concatenated with hidden state of the decoder and then predict.

4.3 Jointly train encoder and decoder models using mini-batches

We built an encoder and decoder recurrent neural network (RNN) with long short-term memory units (LSTM) so that the model can capture word dependencies [15]. The embedding dimension is 300, and the dimensionality of the internal state is set to 512.

4.4 Implement greedy-search decoding module and beam-search decoding

A simple approximation is to use a greedy search that selects the most likely word at each step in the output sequence. This approach has the benefit that it is very fast, but the quality of the final output sequences may be far from optimal.

The beam search that expands upon the greedy search and returns a list of most likely output sequences.Instead of greedily choosing the most likely next step as the sequence is constructed, the beam search expands all possible next steps and keeps the k most likely, where k is a user-specified parameter and controls the number of beams or parallel searches through the sequence of probabilities.

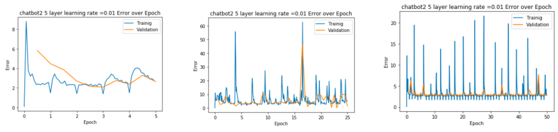
5 Experiment

5.1 Chat bot 1

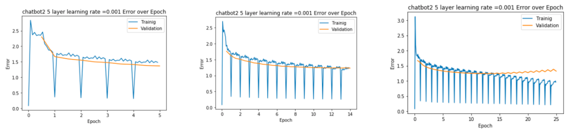
Chat bot 1 is trained on Cornell Movie dataset. In order to decrease the error, we tried two learning rate, 0.01 and 0.001. The result is shown in Fig 5. Apparently, at learning rate 0.001, the training error and validation error can decrease to as low as 1.2.

5.2 Chat bot 2

Chat bot 2 is trained on Daily dialogue dataset. As shown in Fig 3, we conducted our experiment on



(a) learning rate = 0.01



(b) learning rate = 0.001

Figure 3: Learning rate of Chat bot 2

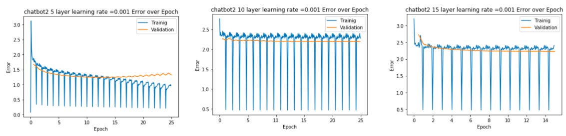


Figure 4: 5,10,15 Hidden layer on chat bot 2

chat bot 2 with learning rate 0.01 and 0.001. For learning rate 0.01, the training reached 50 epoch, the training error and validation error won't decrease with the increase of epoch. For learning rate 0.001, the error can decrease to 1.2 with only 5 epoch,however, the error stable at 2.6 even trained to 50 epoch at learning rate 0.01. We also increased the number of hidden layer to understanding the model, as shown in Fig 4.

5.3 Chat bot 3

Chat bot 3 is trained on the mix of Cornell Movie dataset and Daily Dialogue dataset. We used 0.01 and 0.001 as our learning rate. The learning rate of 0.001 has better performance. In the future, we should explore more learning rate to decrease the error.

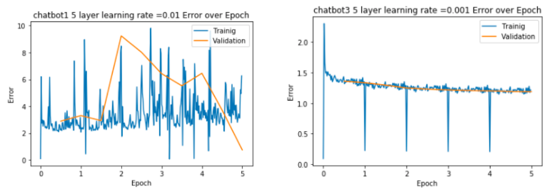


Figure 5: Learning rate 0.01 and 0.001 on chat bot 3

	Grammaticality	Naturalness	Interestingness
Chat bot 1, 5 epoch	3.3	2	1.7
Chat bot 2, 5 epoch	4	2.7	2.7
Chat bot 2, 14 epoch	1	1	1.7
Chat bot 2, 25 epoch	1	1	2.3
Chat bot 3, mixed dataset	3.7	4	2

Table 2: Average score for each criteria

Hi
I am not going to get you to theout.
Where are you from?
I am not going to beout.
What do you do for a living?
I am not going to beout.
Can you let me know what time it is?
I am not going to get out of here.
Do you have time to go shopping tomorrow?
I am not going to get aout.
When is your birthday?
I am not going to get you.
Which one do you prefer, a dog or a cat?
I am not going to get you to theout.
When is the next Olympic?
I am not going to get you.
Can you recommend any country to go on a trip?
I am not going to beout.
How is the weather today?
I am not going to get it out of my life.
I miss my family.
I am not going to beout.

Table 3: the input and output of chat bot trained with Cornell Movie Corpus

Hi
I am glad , I am not sure of you ..
Where are you from?
I am glad ofout ..
What do you do for a living?
I am glad , but I will be able to be aout .
Can you let me know what time it is?
I am sorry , but I am not sure ..
Do you have time to go shopping tomorrow?
I am glad I will ..
When is your birthday?
I am sorry , but I am not sure of theout ..
Which one do you prefer, a dog or a cat?
I am sorry , but I am not sure of you ..
When is the next Olympic?
I am glad , but I am not sure of you ..
Can you recommend any country to go on a trip?
I am glad of you ..
How is the weather today?
I am sorry , but I am not sure of you ..
I miss my family.
I am sorry ..

Table 4: the input and output of chat bot trained with Daily dialogue, 5 epoch. Underfitting result: training error larger than validation error

6 Evaluation

We only conduct human evaluation to the outputs as it has been debated that it is the only measure that open-ended generation tasks can rely on (Li et al., 2016), (Wiseman et al., 2017). Indeed, it has been questioned whether automatic metrics, such as BLEU, are appropriate to capture response quality of open-ended generation tasks (Dai and Le, 2015) (Galley et al., 2015). Considering that open-ended generation does not aim to derive any correct answer, we characterize (1) grammaticality and (2) naturalness (3) interestingness for some samples of the four different models by conducting human evaluation. We asked three people to evaluate each criteria for each model, and average score of each criteria was calculated. Each criteria was evaluated compared to each model.

6.1 Grammaticality

For grammaticality, with a scale of 0-5, we evaluate grammatical errors such as whether a model obeys subject verb agreement, whether a model generates a random string of words or a full sentence, and which kind of tense it can generate. The chat bot 2, where 5 epoch was used, performs the best in terms of grammaticality. Compared to other models, it generates a grammatical full sentence, which obeys subject verb agreement

Hi
conversation sets awful pardon romance brilliant parcel discuss lisa efficient typicalout swimming
How are you?
conversation sets wilson eaten own conversation level typicalout swimming
What do you do for a living?
conversation sets awful romance level pleaseout dance discuss swimming
Can you let me know what time it is?
conversation sets wilson eaten own however level lisa cameras lisa coach western swimming
Do you have time to go shopping tomorrow?
conversation sets wilson eaten own owever level lisa cameras lisa coachout swimming
When is your birthday?
conversation sets awful romance maintain register swimming
Which one do you prefer, a dog or a cat?
conversation sets awful romance maintain register swimming
When is the next Olympic?
conversation sets awful romance maintain register swimming
Can you recommend any country to go on a trip?
conversation sets awful romance maintain register swimming
How is the weather today?
conversation sets awful romance maintain register swimming
I miss my family.
conversation sets awful romance maintain register swimming

Table 5: the input and output of chat bot trained with Daily dialogue, 14 epoch.Just right result: training error approximately equals to validation error

Hi
cancer demand charges songs exciting hong speed.
Where are you from?
cancer demand charges songs magazine palace speed cancer ! cheap santa tend safe haven surpriseout speed.
What do you do for a living?
cancer demand charges songs magazine palace speed cancer ! cheap santa tend safe hospital nice speed.
Can you let me know what time it is?
cancer demand allowed phone independent cancer demand cheap police speed.
Do you have time to go shopping tomorrow?
cancer demand allowed phone independent cancer digital certainly safe towards :: definitelyout speed.
When is your birthday?
cancer demand allowed phone independent cancer digital certainly safe towards mexico surprise library speed.
Which one do you prefer, a dog or a cat?
cancer demand allowed phone independent cancer demand cheap police whom cancer judge speed.
When is the next Olympic?
cancer demand allowed phone independent cancer digital certainly safe towards mexico definitelyout speed.
Can you recommend any country to go on a trip?
cancer demand allowed phone independent cancer digital certainly safe towards mexico surpriseout speed.
How is the weather today?
cancer demand allowed phone independent cancer digital certainly safe towards mexico songs speed.
I miss my family.
cancer demand charges songs certainly surprise wear next speed.

Table 6: the input and output of chat bot trained with Daily dialogue, 25 epoch. Overfitting result: training error less than validation error

and can generate future tense. Even if it generates meaningless words, it happens less than other models. Also, it puts a comma and a punctuation mark at the right place. On the other hand, the chat bot 3 performs similar to the chat bot 2 with 5 epoch, but it misses a punctuation mark between two sentences. Some models, such as chat bot 2 with 25 epoch and with 14 epoch performs not very well as they only generate a random string of words.

Hi
I am not sure I am not going to be aout.
How are you?
I am not sure I am not sure.
What do you do for a living?
I am not sure I am not going to be aout.
Can you let me know what time it is?
I am not sure.
Do you have time to go shopping tomorrow?
I am not sure I am a littleout.
When is your birthday?
I am not sure I am not going to be able to be aout.
Which one do you prefer, a dog or a cat?
I am not sure I am not going to be able to be aout.
When is the next Olympic?
I am not sure.
Can you recommend any country to go on a trip?
I am not sure.
How is the weather today?
I am not sure I am not going to be aout.
I miss my family.
I am not sure.

Table 7: the input and output of chat bot trained with mixed dataset, both Cornell Movie Corpus and Daily dialogue.

6.2 Naturalness

For naturalness, with a scale of 0-5, we evaluate whether a response from a model is similar to natural dialogue. All of the models perform not very well on naturalness as they only repeat either the same string of words or the same sentence. However, the chat bot 3 trained with a mixed dataset was considered as performed the best. This is because for some questions asked to the chat bot, it makes sense to answer with the repetitive sentence that it generates,such as I am not sure.

6.3 Interestingness

For interestingness, with a scale of 0-5, we evaluate whether a response from a chat bot evokes a person to continue talking to it. All of the responses generated from each model was not very interesting to continue talking as they all repeat the same sentence or words.

7 Conclusion and future work

We trained chat bots to produce open-ended generation by changing some hyper-parameters, such as epoch, num layers, and learning rate, and reported the results. The biggest problem of the chat bots was that they repeat the same string of words or a sentence. Thus, in order to understand the model better, we need to conduct more experiments on other parameters, such as batch size, rnn size, learning rate decay, min learning rate, and keep probability.

Acknowledgments

We are thankful to Prof. Marilyn Walker who provided expertise that greatly assisted the research.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In *Advances in neural information processing systems*, pages 3079–3087.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltableu: A discriminative metric for generation tasks with intrinsically diverse targets. *arXiv preprint arXiv:1506.06863*.

Supratip Ghose and Jagat Joyti Barua. 2013. Toward the implementation of a topic specific dialogue based natural language chatbot as an undergraduate advisor. In *2013 International Conference on Informatics, Electronics and Vision (ICIEV)*, pages 1–5. IEEE.

Matthew Henderson. 2015. Machine learning for dialog state tracking: A review.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

Minghui Qiu, Feng-Lin Li, Siyu Wang, Xing Gao, Yan Chen, Weipeng Zhao, Haiqing Chen, Jun Huang, and Wei Chu. 2017. 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.

Silvia Quarteroni and Suresh Manandhar. 2007. A chatbot-based interactive question answering system. *Decalog 2007*, 83.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing*, pages 583–593. Association for Computational Linguistics.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*.

Yiping Song, Rui Yan, Xiang Li, Dongyan Zhao, and Ming Zhang. 2016. Two are better than one: An ensemble of retrieval-and generation-based dialog systems. *arXiv preprint arXiv:1610.07149*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Zhuoran Wang and Oliver Lemon. 2013. A simple and generic belief tracking mechanism for the dialog state tracking challenge: On the believability of observed information. In *Proceedings of the SIGDIAL 2013 Conference*, pages 423–432.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

Jason D Williams and Geoffrey Zweig. 2016. End-to-end lstm-based dialog control optimized with supervised and reinforcement learning. *arXiv preprint arXiv:1606.01269*.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052*.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2016. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint arXiv:1612.01627*.