Improving open domain Dialogue systems

Nitin Kishore Sai Samala
College of Information and Computer Sciences
University of Massachusetts Amherst
140 Governors Drive, Amherst, MA 01003
nsamala@cs.umass.edu

Abstract

Conversational modeling is an important task since it can be viewed as a cognitive system, which must carry out natural language understanding, reasoning, decision making and natural language generation to replicate or emulate the behavior of the agents in the training corpus. Retrieval-based models use a repository of predefined responses restricted to specific domains and a heuristic to pick an appropriate response based on the input and context while generative models produce system responses that are autonomously generated word-by-word, opening the possibility for realistic, flexible interactions and can be trained end-to-end. Attention and Intention play intrinsic roles in any conversation or a dialogue process. As expected the lack of consistency is a common failure mode of the Sequence to sequence models that extract knowledge and perform simple forms of common sense reasoning on large general domain dataset of movie subtitles. The responses generated tend to be safe, commonplace responses (e.g., I dont know) regardless of the input because the traditional objective function, i.e., the likelihood of output (response) given input (message) is unsuited to response generation tasks. Such issues hold the chatbots back from being ready for effective application in the industry. In this project, I tackle these issues by combining an attention mechanism with a diversity promoting objective function, MMI (Maximum mutual information) to generate more interesting responses that are relevant and not likely to be picked by MLE and have higher lexical as well as sentential diversity than baseline models and generates more acceptable diverse output yielding satisfactory results in human evaluation.

1. Introduction

Natural language conversation is one of the most challenging artificial intelligence problems, which involves language understanding, reasoning, and the utilization of common sense knowledge. As conversational agents gain traction in user interfaces, facilitating smooth interaction between humans and their electronic devices, yet continuing to face major challenges in the form of robustness, scalability and domain adaptation, there has been growing research interest in training naturalistic conversation systems from large volumes of human-to-human interactions. Companies start off by outsourcing their conversations to human workers and promise of automation once enough data has been collected. A chatbot should be purposeful, reflective of the products voice, and sympathetic with the users. The scripts tone can be familiar or professional. The bot can be polite and conversational or entirely focused on the task at hand. Given that users innately treat computers as social beings, theres no need to pretend to be a human. Instead, a bot is a potential opportunity to expand the corporate voice to the familiar.

Many companies hoping to develop bots to have natural conversations indistinguishable from human ones are using NLP and Deep Learning techniques to make this possible. Architectures like sequence-to-sequence are uniquely suited for generating text and researchers are hoping to make rapid progress in this area [13]. Retrieval-based models use a repository of predefined responses and a heuristic to pick an appropriate response based on the input and context while Generative models create new responses from scratch. A retrieval-based open domain system is obviously impossible because you can never handcraft enough responses to cover all possible queries [10]. A generative open-domain system is almost Artificial General Intelligence (AGI) because it needs to handle all possible scenarios. Until recently, the goal of training open-domain conversational systems that emulate human conversation has seemed elusive but vast quantities of conversational exchanges now available raise the prospect of building data-driven models that can begin to communicate conversationally [6]. Were still in a nascent phase of building generative models that work reasonably well so production systems are more likely to be retrieval-based for now. There are few challenges, most of which being active research areas, when building conversa-

tional agents. The following three are the most prominent.

One major issue for these data-driven systems is their propensity to select the response with greatest likelihoodin effect a consensus response of the humans represented in the training data. Outputs are frequently vague or non-committal (Li et al., 2016), and when not, they can be wildly inconsistent. Lack of Intention and Diversity, resulting in generic responses like Thats great! or I dont know that work for a lot of input cases is an ubiquitous issue in MLE decoding giving credence to the hypothesis that MLE is not a suitable objective function for language generation. To get non-boring outputs, we need to urge the model to pick less probable responses. In other words, choose a diversity promoting objective function. [7]

A conversation, the communication of thoughts through words is a structural process intimately connected with two nonlinguistic notions: intention and attention. Attention explicates the processing of utterances, for example, paying attention to specific words in a sentence, while intention has its primary role of explaining discourse structure and coherence [15], [12]. Perhaps the most common and successful approach has been to view the dialogue problem as a partially observable Markov decision process. Not incorporating Context, or basically keeping track of what has been said and what information has been exchanged to produce more sensible responses would give the model a severe handicap. The output would be a series of highly probable n-grams that co-occur frequently and have nothing to do with the query.

Another issue is that, natural dialogue is not deterministic; for example, the replies to Whats your name and where do you come from? will vary from person to person [2]. Dialog systems learn to generate linguistic plausible responses, but arent trained to generate semantically consistent answers to identical inputs. (Vinyals and Le, 2015) suggests that the lack of a coherent personality makes it impossible for current systems to pass the Turing test. Li et al. (2016b) have proposed learning representations of personas to account for interpersonal variation, but there can be variation even among a single persons responses to certain questions. We can endow data-driven systems with the coherent persona needed to model human-like behavior. [8]

This project investigates the task of building open domain, generative conversational dialogue systems on large dialogue corpora using Maximum Mutual Information (MMI) as the objective function in neural models to produce more diverse, interesting, and appropriate responses, yielding satisfactory results in human evaluations while combining it with an attention mechanism to make sure the diverse responses are relevant to the context. This project culminates an implementation, combining two of these methods to build a better conversational model that has higher lexical as well as sentential diversity than baseline models and generates more acceptable diverse output than sampling from a deterministic decoder.

2. Related work

A conversation process may be cast as a sequence-to-sequence mapping task that stands in contrast to conventional dialog systems, which typically are template or heuristic driven even where there is a statistical component. Neural network based approaches have been successfully applied in sequence-to-sequence mapping tasks and have made significant progresses in machine translation, language understanding and speech recognition [14], [13]. To learn conversational patterns from data: researchers have begun to explore data-driven generation of conversational responses within the framework of statistical machine translation (SMT), either phrase-based (Ritter et al.,2011), or using neural networks to re-rank, or directly in the form of sequence-to-sequence (SEQ2SEQ) models that requires little feature engineering and domain specificity. Conversational modeling can directly benefit from this formulation because it requires mapping between queries and responses. From a qualitative point of view, the model is sometimes able to produce natural conversations. Among these neural network-based approaches, one approach, which is called encoder-decoder framework, aims at relaxing much requirement on human labeling [3].

A persona can be viewed as a composite of elements of identity (background facts or user profile), language behavior, and interaction style. It is also adaptive, since an agent may need to present different facets to different human interlocutors depending on the interaction. These persona vectors are trained on human-human conversation data and used at test time to generate personalized responses. (Li et al., 2016) experiments on an open-domain corpus of Twitter conversations and dialog datasets comprising TV series scripts show that leveraging persona vectors can improve relative performance up to 20% in BLEU score and 12% in perplexity, with a commensurate gain in consistency as judged by human annotators. Recently, Serban et al. (2017) have introduced latent variables to the dialogue modelling framework, to model the underlying distribution over possible responses directly. At generation time, we can sample a response from the distribution by first sampling an assignment of the latent variables, and then decoding deterministically. They introduce stochasticity without resorting to sampling from the decoder, which can lead to incoherent output. [11]

An engaging response generation system should be able to output grammatical, coherent responses that are diverse and interesting. In addition, shorter responses typically have higher likelihoods, and so wide beam sizes often result in very short responses (Tu et al., 2017; Belz, 2007). To resolve this problem, Li et al. (2016a) propose instead us-

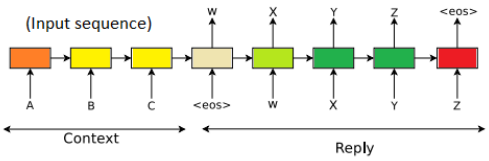

Figure 1. Sequence to sequence model

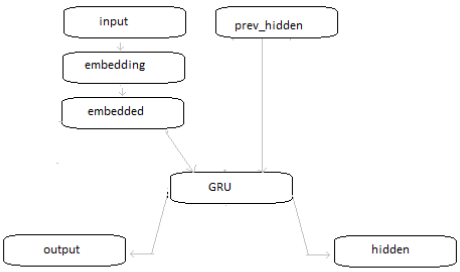

Figure 2. Encoder network

ing Maximum Mutual Information (MMI), first introduced in speech recognition (Bahl et al., 1986; Brown, 1987), as an optimization objective that measures the mutual dependence between inputs and outputs, with a length boost as a decoding objective, and report more interesting generated responses.

Prior work in generation has sought to increase diversity, but with different goals and techniques.Carbonell and Goldstein (1998) and Gimpel et al.(2013) produce multiple outputs that are mutually diverse, either non-redundant summary sentences or N-best lists. My goal, however, is to produce a single non-trivial output while using attention and this method does not require identifying lexical overlap to foster diversity

3. Methodology

The sequence to sequence architecture, where two recurrent neural networks work together to transform one sequence to another as shown in Fig1. has an encoder network compresses an input sequence into a vector, and a decoder network which unfolds that vector into a new sequence [13].Unlike sequence prediction with a single RNN, where every input corresponds to an output, the seq2seq model frees us from sequence length and order.

The encoder of a seq2seq network is a RNN that outputs some value for every word from the input sentence. For every input word the encoder outputs a vector and a hidden state, and uses the hidden state for the next input word. With a seq2seq model the encoder creates a single vector which, in the ideal case, encodes the meaning of the input sequence into a single vector a single point in some N dimensional space of sentences. For every input word, the encoder outputs a vector and a hidden state, and uses the hidden state for the next input word.This context vector is used as the initial hidden state of the decoder. [1]

At every step of decoding, the decoder is given an input token and hidden state. The initial input token is the start-of-string SOS token, and the first hidden state is the context vector (the encoders last hidden state).

3.1. Model - Attention-MMI

Attention allows the decoder network to focus on a different part of the encoders outputs for every step of the decoders own outputs. If only the context vector is passed between the encoder and decoder, that single vector will carry the burden of encoding the entire sentence. The encoder output vectors are multiplied by the calculated attention weights using a feed forward network using the decoders input and hidden state as inputs to create a weighted combination. The result called attn applied in the code should contain information about that specific part of the input sequence, and thus help the decoder choose the right output words.

Since there are sentences of variable size in the training data, to actually create and train the attn layer we have to choose a maximum sentence length which is the input length, for encoder outputs that it can apply to. Sentences of the maximum length will use all the attention weights, while shorter sentences will only use the first few. For this architecture I am using a Gated Recurrent unit instead of LSTM. The idea behind a GRU layer is quite similar to that of a LSTM layer, as are the equations. [1]

$$z = \sigma(x_t U^z + s_{t-1} W^z)$$
$$r = \sigma(x_t U^r + s_{t-1} W^r)$$
$$h = \tanh(x_t U^h + (s_{t-1} \circ r) W^h)$$
$$s_t = (1 - z) \circ h + z \circ s_{t-1}$$

A GRU has two gates, a reset gate r, and an update gate z. Intuitively, the reset gate determines how to combine the new input with the previous memory,s, at time step t, and the update gate defines how much of the previous memory to keep around. If we set the reset to all 1s and update gate to all 0s we again arrive at our plain RNN model.The input and forget gates are coupled by an update gate z and the reset gate r is applied directly to the previous hidden state. Thus, the responsibility of the reset gate in a LSTM is really split up into both r and z. [4]

3.2. Objective function

To deal with the issue that SEQ2SEQ models tend to generate generic and commonplace responses such as "I dont know", we follow Li et al.(2016) by re-ranking the generated N-best list using a scoring

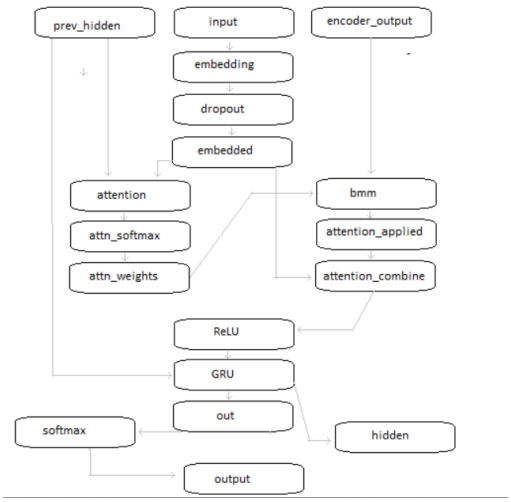

Figure 3. Attention decoder network

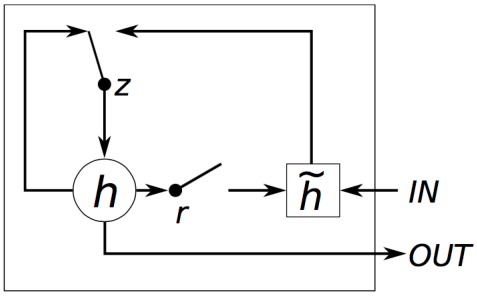

Figure 4. GRU Gating. Chung, Junyoung, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling. (2014)

function. Let S denote an input message sequence (source) $\mathcal{S} = \{s_1, s_2, s_3, \ldots, s_{N_s}\}$ and target $\mathcal{T} = \{t_1, t_2, t_3, \ldots, t_{N_t}, EOS\}$ The standard objective function is the log-likelihood of target T given source S, which at test time yields the statistical decision problem that only selects for targets given sources, not the converse.

$\widehat{T} = argmax_T \{log(p(T|S))\}$

In MMI, parameters are chosen to maximize (pairwise) mutual information between the source S and the target T $log\frac{p(S,T)}{p(S)p(T)}$ which gives us the MMI objective function as :

$\widehat{T} = argmax_T \{log(p(T|S)) - \lambda log(p(T))\}$

where the first term denotes the probability of the generated response given the message and λ denotes the associated penalty weight. A threshold for number of words γ is chosen as well. To compute p(S—T), we need to train an inverse SEQ2SEQ model by swapping messages and responses. We replace the language model p(T) with U(T),which adapts the standard language model by multiplying by a weight g(k) that is decremented monotonically as the index of the current token k increases:

$U(T) = \prod_{i=1}^{N_t} p(t_k|t_1, t_2, t_3, \ldots, t_{k-1}).g(k)$

If k is less than γ then we set g(k) as 1 else 0. As the influence of the input on decoding declines, the influence of the language model comes to dominate. [7]

Thus the final objective function for which direct decoding is tractable becomes

$\widehat{T} = argmax_T \{log(p(T|S)) - \lambda log(U(T))\}$

We can optimize γ and λ on N-best lists of response candidates generated from the development set using MERT (Och, 2003) by optimizing BLEU.I have chosen the values mentioned in the paper [7] and tuned it a bit since it is the same dataset. The Kullback-Leibler divergence resembles minimizing mutual information and maximizing likelihood.

4. Datasets

For this project,I am using Open subtitles dataset and Cornell Movie-Quotes corpus. The motivation behind choosing these is due to the wide range of context in the conversation between characters which models the open domain dialog closely.

4.1. Cornell Movie-Quotes Corpus

This corpus contains a large metadata-rich collection of fictional conversations extracted from raw movie scripts:
- 220,579 conversational exchanges between 10,292 pairs of movie characters
- involves 9,035 characters from 617 movies
- in total 304,713 utterances
Movie metadata included:
- genres
- release year
- IMDB rating
- number of IMDB votes
- IMDB rating
Character metadata included:
- gender (for 3,774 characters)
- position on movie credits (3,321 characters)

4.2. Open Subtitles

Open Subtitles dataset (Tiedemann, 2009). This dataset consists of movie conversations in XML format. It contains sentences uttered by characters in movies. I applied a simple processing step removing XML tags and obvious non-conversational text (e.g., hyperlinks) from the dataset. As turn taking is not clearly indicated, I treated consecutive sentences assuming they were uttered by different characters. The model needs to predict the next sentence given the previous one, and do did this for every sentence (noting that this doubles the dataset size, as each sentence is used

Category	OpenSubs	Cornell
Train pairs	14008024	108135
Trimmed pairs	912334	4056
Question vocab	65015	3468
Response vocab	74592	3846
Validation pairs	10000	30000
Trimmed pairs	628	1117
Question vocab	851	1686
Response vocab	1069	1676

Table 1. Corpus statistics.

both for context and as target). The training and validation split has 62M sentences (923M tokens) as training examples, and the validation set has 26M sentences (395M tokens). The split is done in such a way that each sentence in a pair of sentences either appear 228 together in the training set or test set but not both. Open Subtitles is quite large, and rather noisy because consecutive sentences may be uttered by the same character.

4.3. Pre-processing

Preprocessing involved tokenization, lower case normalization, regular expressions to deal with contraction words and filtering pairs that exceed the max sequence length after appending the ¡EOS¿ token. The maximum sequence of 20 tokens was chosen for this project.From the raw data I extracted all the questions and answers with following constraints:
- question should end with "?"
- answer is declarative
- answer is less than 20s after the question (Since the data here is basically subtitles of conversations between characters in TV shows and movies)

Examples from Opensubs:
['is she ?', 'she s really a genius']
['who is feeling grassy ?', 'you are']

Examples from Cornell:
['game six is history pal', 'you re not making sense']
['yes . i told you i was your number one fan', 'i m getting to believe you']

5. Experiments and results

I didnt use pre-trained word vectors in my experiments, but adding an embedding layer (the matrix E in our code) makes it easy to plug them in. The embedding matrix is really just a lookup table the ith column vector corresponds to the ith word in our vocabulary. By updating the matrix E we are learning word vectors ourselves, but they are very specific to our task (and data set) and not as general as those that you can download, which are trained on millions or billions of documents.Compared to the dozens of characters that might exist in a language, there are many many more words, so the encoding vector is much larger.

5.1. Training details

There is no principled reason why Ive chosen GRUs instead LSTMs in this part. GRUs are quite new (2014), and their tradeoffs havent been fully explored yet. According to empirical evaluations in [4] and [5], there isnt a clear winner. In many tasks both architectures yield comparable performance and tuning hyperparameters like layer size is probably more important than picking the ideal architecture. GRUs have fewer parameters (U and W are smaller) and thus may train a bit faster or need less data to generalize. To optimize RNN performance instead of learning from one sentence at a time, I grouped sentences of the same length (even padded all sentences to have the same length of 20) and then perform large matrix multiplications and sum up gradients for the whole batch. I ran the regular seq2seq model and the modified attention-MMI models on NVIDIA Titan X GPUs on the gypsum cluster using cuda porting in pytorch. The model was executed for 590,000 iterations or 7 epochs on Opensubs and 100,000 iterations or 10 epochs on Cornell dataset, using Adam for optimization. I have also used teacher forcing of 0.5. Teacher forcing is the concept of using the real target outputs as each next input, instead of using the decoders guess as the next input.

5.2. Code

. Please refer to the link for the github repository in the footnotes.[1]

5.3. Inferences

The model trained without the MMI objective function yielded "I m gonna go EOS" and "I m fine EOS" as responses to several inputs from Opensubs dataset which weren't relevant in most of the cases. This goes to demonstrate that MLE decoding favoring responses that unconditionally enjoy high probability, and instead biases towards those responses that are specific to the given input which is the problem statement. Using teacher forcing helped it to converge faster but when the trained network is exploited, it exhibits some instability. Irrelevance in the some answers might also be attributed to the training data. The corpus is filled with conversations that have a prior context and certain responses make sense between those characters so the generated responses will be based on the sequences observed from these subtitles.

The last result from table 2 shows attention mechanism at work. We see that Attention-MMI generates significantly

[1]Code uploaded on https://github.com/snknitin/ChatBot-Text-Summarizer

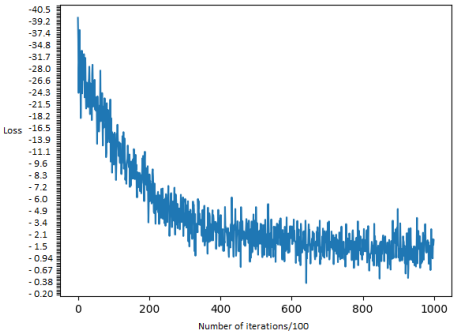

Figure 5. Loss plot for Cornell dataset

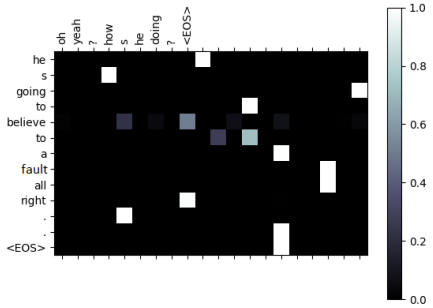

Figure 6. Visualizing attention on a sentence evaluation from valid set of Cornell

more interesting outputs than SEQ2SEQ. we can imagine looking where the network is focused most at each time step, in this case the word "tesla".Figures 7 and 8 show attention output displayed as a matrix, with the columns being input steps and rows being output steps.It is interesting to look at the semantic dependencies of these sentences over multiple time steps. Without attention, the MMI decoding would have just yielded an assemblage of tokens that have a low probability of being selected.Attention gives relevance and context.

I also observed that ungrammatical segments tend to appear in the later parts of the sentences, especially in long sentences.The first words to be predicted significantly determine the remainder of the sentence. This is an expected phenomenon from the diversity promoting objective function that i modified.

5.4. Illustrations and graphs

Please refer to loss plots and attention visualizations in Figs 5,6,7.

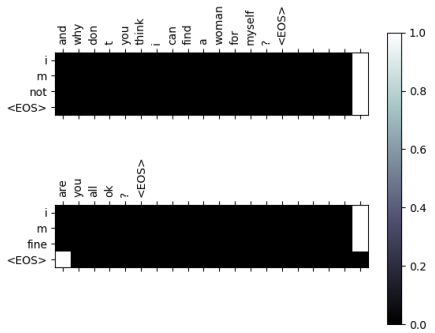

Figure 7. Visualizing attention on a sentences evaluated from valid set of Opensubs

6. Evaluation

The ideal way to evaluate a conversational agent is to measure whether it is fulfilling its task, e.g. solve a customer support problem, in each conversation but such labels are expensive to obtain because they require human judgment and evaluation. Sometimes there is no well-defined goal, as is the case with open-domain models.

Common metrics such as BLEU that are used for Machine Translation and are based on text matching arent well suited because sensible responses can contain completely different words or phrases. Researchers find that none of the commonly used metrics really correlate with human judgment or give a measure of how good the model is [9]. Which is why instead of an evalution metric like BLEU or Rouge, i chose to perform a small human evaluation.

I had a diverse group of 5 volunteers grade responses to specific set of inputs. I randomly sampled 40 question pairs from both datasets into multiple choice questions, varying the option choices with a generic boring response and attention-MMI response alternating each time and asked to select a more relevant answer that seems meaningful and interesting, ignoring grammar constraints. The responses from attention-MMI were picked 28 times on average out of 40. Apart from this, a qualitative analysis of the responses should prove to be a better standard to judge the quality of the model.

7. Conclusions and Future work

Even with grammatical inconsistencies that is expected in generative models creating responses from scratch , the resulting predictions show surprising relevance to the question context while being very diverse and non-generic. Instead of just predicting the most probable sequence, either using greedy or beam search decoding, which mostly resulted in Im gonna go ,Im fine or I dont know as the top replies in MLE estimation, MMI attention model resulted

Input	Response OpenSub	Response Cornell
what is your name ?	that is it so needing	none behind i kept was an eye
when were you born ?	well most likely not really capable of it	gee if we have a good person
are you a leader or a follower ?	im not now	theres nothing wrong about that
see you later	yes a humble to me	gee that figures were no longer
what is the purpose of living ?	oh tell her to make you you lose my home	impossible i have no intention but theyre re
what is the purpose of being intelligent ?	if someones you	theyve decided about this secret
what is moral ?	the bullet in the palace its is a common profession	nonsense theres no one of the american
what is morality ?	i dont know the right	yah im impressed to remember
what is your job ?	visiting getting the back out	revenge i wanted her to say
what do you like to talk about?	i UNK holly	maybe we come out there this thing
what do you think about tesla?	ive accomplished of america who kill the power	thats not crazy in stock

Table 2. Results. Predicted sequences from Attention-MMI model

in much more engaging and interesting replies that were not necessarily the most probable choice but were consistent with the question instead of a random assemblage of words. This concurs with the hypothesis that MLE is not a suitable objective function for natural dialogue and incorporating attention mechanism during decoding helped maintain the context of the question and helped generate meaningful responses. Intermingling these two concepts improved the performance of the model in an open domain query without restrictions in replies or any hard-coded responses and so corroborates the inadequacy of perplexity as an evaluation metric for dialogue models (Liu et al., 2016).

Future work in this project could include further optimization of the model hyperparameters, observing the effect of feeding in the sequences in reverse to the GRU layer, using LSTMs for attention in longer sequences, experimenting with other objective functions like including the poisson distribution of the target sequence in the objective function, incorporating a persona in the model to have consistent replies and maybe hard-coding certain rules and responses to frequent queries to have them be more grammatical.There are other forms of attention that work around the length limitation by using a relative position approach which could offer some interesting avenues to pursue.There appears to be a Goldilocks region of the probability space, where the responses are interesting and coherent. Finding ways of concentrating model samples to this region is thus a potentially promising area of research for open-domain dialogue agents.

References

[1] PyTorch translation with a sequence to sequence network and attention. `http://pytorch.org/tutorials/_sources/intermediate/seq2seq_translation_tutorial.rst.txt`. Accessed: 2010-09-30.

[2] K. Cao and S. Clark. Latent variable dialogue models and their diversity. *arXiv preprint arXiv:1702.05962*, 2017.

[3] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[4] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[5] R. Jozefowicz, W. Zaremba, and I. Sutskever. An empirical exploration of recurrent network architectures. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2342–2350, 2015.

[6] R. Kadlec, M. Schmid, and J. Kleindienst. Improved deep learning baselines for ubuntu corpus dialogs. *arXiv preprint arXiv:1510.03753*, 2015.

[7] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

[8] J. Li, M. Galley, C. Brockett, G. P. Spithourakis, J. Gao, and B. Dolan. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*, 2016.

[9] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*, 2016.

[10] R. Lowe, N. Pow, I. Serban, and J. Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*, 2015.

[11] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. C. Courville, and Y. Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301, 2017.

[12] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*, 2015.

[13] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[14] O. Vinyals and Q. Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

[15] K. Yao, G. Zweig, and B. Peng. Attention with intention for a neural network conversation model. *arXiv preprint arXiv:1510.08565*, 2015.