Why are Sequence-to-Sequence Models So Dull?

Understanding the Low-Diversity Problem of Chatbots

Shaojie Jiang
University of Amsterdam
Amsterdam, The Netherlands
s.jiang@uva.nl

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

Abstract

Diversity is a long-studied topic in information retrieval that usually refers to the requirement that retrieved results should be non-repetitive and cover different aspects. In a conversational setting, an additional dimension of diversity matters: an engaging response generation system should be able to output responses that are diverse and interesting. Sequence-to-sequence (Seq2Seq) models have been shown to be very effective for response generation. However, dialogue responses generated by Seq2Seq models tend to have low diversity. In this paper, we review known sources and existing approaches to this low-diversity problem. We also identify a source of low diversity that has been little studied so far, namely model over-confidence. We sketch several directions for tackling model over-confidence and, hence, the low-diversity problem, including confidence penalties and label smoothing.

1 Introduction

Sequence-to-sequence (Seq2Seq) models (Sutskever et al., 2014) have been designed for sequence learning. Generally, a Seq2Seq model consists of two recurrent neural networks (RNN) as its encoder and decoder, respectively, through which the model cannot only deal with inputs and outputs with variable lengths separately, but also be trained end-to-end. Seq2Seq models can use different settings for the encoder and decoder networks, such as the number of input/output units, ways of stacking layers, dictionary, etc. After showing promising results in machine translation (MT) tasks (Sutskever et al., 2014; Wu et al., 2016), Seq2Seq models also proved to be effective for tasks like question answering (Yin et al., 2015), dialogue response generation (Vinyals and Le, 2015), text summarization (Nallapati et al., 2016), constituency parsing (Vinyals et al., 2015a), image captioning (Vinyals et al., 2015b), and so on.

Seq2Seq models form the cornerstone of modern response generation models (Vinyals and Le, 2015; Li et al., 2015; Serban et al., 2016, 2017; Zhao et al., 2017). Although Seq2Seq models can generate grammatical and fluent responses, it has also been reported that the corpus-level diversity of Seq2Seq models is usually low, as many responses are trivial or non-committal, like "I don't know", "I'm sorry" or "I'm OK" (Vinyals and Le, 2015; Sordoni et al., 2015; Serban et al., 2016; Li et al., 2015). We refer to this problem as the *low-diversity* problem.

In recent years, there have been several types of approach to diagnosing and addressing the low-diversity problem. The purpose of this paper is to understand the low-diversity problem, to understand what diagnoses and solutions have been proposed so far, and to explore possible new approaches. We first review the theory of Seq2Seq models, then we give an overview of known causes and existing solutions to the low-diversity problem. We then connect the low-diversity problem to the concept of *model over-confidence*, and propose approaches to address the over-confidence problem and, hence, the low-diversity problem.

2 Sequence-to-Sequence Response Generation

Consider a dataset of message-response pairs (X, Y), where $X = (x_1, x_2, \ldots, x_{|X|})$ and $Y = (y_1, y_2, \ldots, y_{|Y|})$ are the input and output sequences, respectively. During training, the goal is to learn the relationships between X and Y, which can be formulated as maximizing the Seq2Seq model probability of Y given X:

$$\max p(Y|X) = \max \prod_{t=1}^{|Y|} p(y_t|y_{<t}, X), \qquad (1)$$

where $y_{<t} = (y_1, y_2, \ldots, y_{t-1})$ are the ground-truth tokens of previous steps.

Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd Int'l Workshop on Search-Oriented Conversational AI, pages 81–86
Brussels, Belgium, October 31, 2018.
ISBN 978-1-948087-75-9

Usually, Seq2Seq models employ Long Short-Term Memory (LSTM) networks as their encoder and decoder. The way a Seq2Seq models realizes (1), is to process the training inputs and outputs separately. On the encoder side, the input sequence X is encoded step-by-step, e.g., at step t:

$$h_t^{enc} = f_\theta^{enc}(h_{t-1}^{enc}, x_t), \tag{2}$$

where $h_0^{enc} = \mathbf{0}$ is the initial hidden state of the encoder LSTM, and θ is the model parameter. The hidden state of the last step $h_{|X|}^{enc}$ is the vector representation of input sequence X.

Then, the decoder LSTM is initialized by $h_0^{dec} = h_{|X|}^{enc}$ so that output tokens can be based on the input:

$$h_t^{dec} = f_\theta^{dec}(h_{t-1}^{dec}, y_{t-1}), \tag{3}$$

with y_0 as a special token (e.g., _START_) to indicate the decoder to start generation, and y_{t-1} as the ground truth token of the last time step. The hidden state h_t^{dec} is further used to predict the output distribution by using a multi-layer perceptron (MLP) and softmax function:

$$P(y_t|y_{<t}, X) = \frac{\exp(c_i f_\theta^{MLP}(h_t^{dec}))}{\sum_{j=1}^{N} \exp(c_j f_\theta^{MLP}(h_t^{dec}))}, \tag{4}$$

where c_* are possible candidates of y_t, which are usually represented as word embeddings. After obtaining this distribution, we can calculate the loss compared with the ground-truth distribution by using, e.g., the cross-entropy loss function, and then we can back-propagate the loss to force the Seq2Seq model to maximize (1).

At test time at t, the step-wise decoder output distribution is conditioned on the actual model outputs $\hat{y}_{<t}$ and X, and the token with the highest probability is chosen as the output:

$$\hat{y}_t = \arg\max_{y_t} p(y_t|\hat{y}_{<t}, X), \tag{5}$$

which is known as the maximum *a posteriori* (MAP) objective function.

3 Diagnosing the Low-Diversity Problem

In the literature, three dominant viewpoints on the low-diversity problem have been shared: lack of variability, improper objective function, and weak conditional signal. Below, we review these diagnoses of the low-diversity problem, with corresponding solutions, and we add a fourth diagnosis: model over-confidence.

3.1 Lack of variability

Serban et al. (2017); Zhao et al. (2017) trace the cause of the low-diversity problem in Seq2Seq models back to the lack of model variability. The variability of Seq2Seq models is different from that of retrieval-based chatbots (Fedorenko et al., 2017): in this study, we focus on the lack of variability of system responses, while in (Fedorenko et al., 2017), the authors deal with the low variability between responses and contexts.

To increase variability, Serban et al. (2017); Zhao et al. (2017) propose to introduce variational autoencoders (VAEs) to Seq2Seq models. At generation time, the latent variable z brought by a VAE is used as a conditional signal of the decoder LSTM (Serban et al., 2017):

$$h_t^{dec} = f_\theta^{dec}(h_{t-1}^{dec}, y_{t-1}, z), \tag{6}$$

where we omit the contextual hidden states for simplicity.

At test time, z is *randomly* sampled from a prior distribution. Although being effective, the improvement in the degree of diversity of generated responses brought by this kind of method is actually brought by the randomness of z. The underlying Seq2Seq model remains sub-optimal in terms of diversity.

3.2 Improper objective function

Li et al. (2015) notice that the MAP objective function may be the cause of the low-diversity problem, since it can favor certain responses by only maximizing $p(Y|X)$. Therefore, they propose to maximize the mutual information between X, Y pairs:

$$\log \frac{p(X,Y)}{p(X)p(Y)}. \tag{7}$$

With the help of Bayes' theorem, they derive two Maximum Mutual Information (MMI) objective functions:

$$\begin{aligned}\hat{Y} = \arg\max_Y \{ &\log p(Y|X) - \\ &\lambda \log p(Y) + \gamma|Y|\},\end{aligned} \tag{8}$$

and

$$\begin{aligned}\hat{Y} = \arg\max_Y \{ &(1-\lambda) \log p(Y|X) + \\ &\lambda \log p(X|Y) + \gamma|Y|\},\end{aligned} \tag{9}$$

where λ and γ are hyper-parameters. Here, $\log p(Y)$ and $\log p(X|Y)$ are the language model

and a reverse model, respectively, with the latter trained using response-message pairs: (Y, X). Besides the time needed for training a reverse model, it should be noted that both objective functions need the length $|Y|$ of candidate responses, which are maintained in N-best lists generated by beam search. To obtain N-best lists with enough diversity, Li et al. (2015) use a beam size of 200 during testing, which is much more time-consuming than the basic Seq2Seq model.

Influenced by the MMI methods, several beam search based approaches (Li et al., 2016; Vijayakumar et al., 2016; Shao et al., 2017) focus on improving the diversity of N-best lists, in the hope of further enhancing the one-best response diversity. However, there are other faster approaches to the low-diversity problem without using beam search, such as the attention-based model that we describe below.

3.3 Weak conditional signal

Since attention layers (Bahdanau et al., 2014) have been introduced into Seq2Seq models for the MT task, they have also been a *de facto* standard module of Seq2Seq models for response generation. The purpose of Seq2Seq attention layers is different from the purpose of the Transformer model (Vaswani et al., 2017). Transformer proposes to rely only on self-attention and avoid using reccurence or convolutions, while attention layers of Seq2Seq aim at strengthening the input signal.

Although the introducing of attention layers can bring improvements to the response generation task, Tao et al. (2018) argue that the original attention signal often focuses on particular parts of the input sequence, which is not strong enough for the Seq2Seq model to generate specific responses, thus causing the low-diversity problem. The authors propose to use multiple attention heads to encourage the model to focus on various aspects of the input, by mapping encoder hidden states to K different semantic spaces:

$$h_{t,k}^{enc} = W_p^k \cdot h_t^{enc}, \tag{10}$$

where $W_p^k \in \mathbb{R}^{d \times d}$ is a learnable projection matrix. The net effect of the extended attention mechanism is, indeed, improvements in the diversity of generated responses. Readers are referred to (Tao et al., 2018) for more details.

3.4 Model over-confidence

As indicated by Hinton et al. (2015), one can think of the knowledge captured in conversation modeling as a mapping from input sequence X to output sequence Y, i.e., the distribution $P(Y|X)$. Therefore, if responses have a low degree of diversity, the learned distribution $P(Y|X)$ is questionable, as re-confirmed by Li et al. (2015). According to (1), the sequence-level distribution $P(Y|X)$ has a direct relationship with the token-level distribution. Therefore, we hypothesize that the token-level distribution $P(y_t|y_{<t}, X)$, produced at the decoder side, may be the culprit.

The decoder LSTM serves as an RNN language model (RNNLM) conditioned on the input sequence (Sutskever et al., 2014). With time steps increasing, the influence of the input sequence X will become weaker according to (3), and if the token-level distribution $P(y_t|y_{<t}, X)$ is problematic, it will have further effects on subsequent outputs (a "snowball effect"). An attention mechanism (Bahdanau et al., 2014; Tao et al., 2018) can be used to reinforce the influence of the input sequence, but there are still chances that the detrimental effect of $P(y_t|y_{<t}, X)$ is stronger than the input signal.

To analyze the problem of $P(y_t|y_{<t}, X)$, we train a Seq2Seq model[1] without attention layer, and plot the token-level distribution of generic responses in Figure 1. Interestingly, we find that the distributions shown signs of model over-confidence (Pereyra et al., 2017). When an attention mechanism is used, similar distributions can still be observed, as illustrated in Figure 2. From these two figures, we can see a common trend of growing confidence: the highest probabilities at each step keep growing, which confirms our conjecture of a snowball effect. Due to this effect, the final several tokens are of low quality, e.g., the no-attention model in Figure 1 starts to repeat itself, and the word "overlapping" in the attention model in Figure 2 is irrelevant for the user input.

A prediction is confident if the entropy of the output distribution is low. *Over-confidence* is often a symptom of over-fitting (Szegedy et al., 2016), which suggests that the inputs or outputs share much similarity from unknown aspects. Although it is hard to figure out what causes the over-fitting, maximizing entropy can usually help to regularize the model, making it generalize better. In (Pereyra et al., 2017), the authors propose to add the negative

[1] We are using ParlAI framework (Miller et al., 2017).

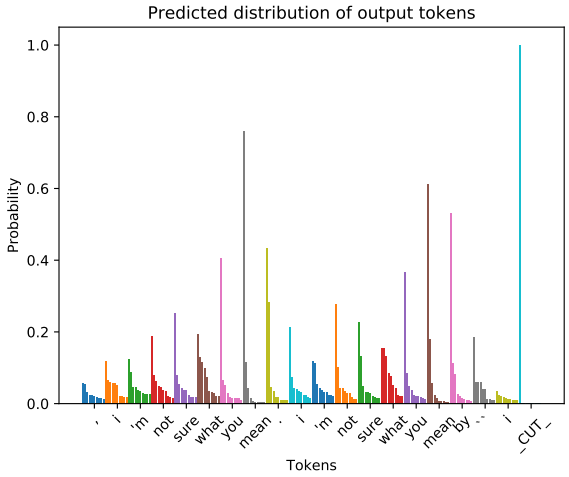

Figure 1: Given the input sequence: *how about we recognize the brilliance in everyone, or in mankind as a whole.*, the predicted distribution of model outputs, and tokens on x axis are MAP predictions. Note that we kept top-10 probabilities at each prediction step for simplicity and this output was cut before the _EOS_ token was emitted.

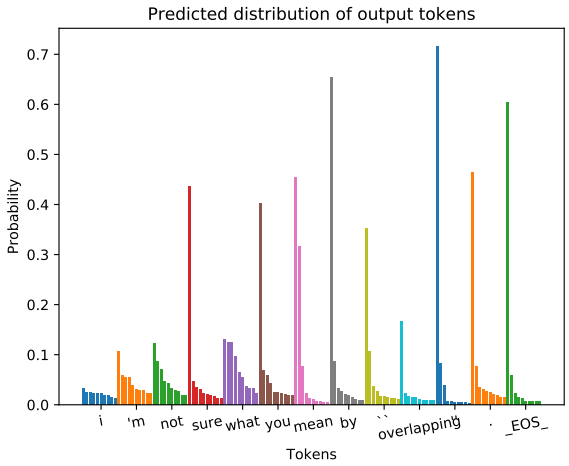

Figure 2: Predicted distribution of the same input as in Figure 1 when an attention mechanism is used.

entropy to the negative log-likelihood loss function during training, which can easily be tailored for conversation modeling:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} \log p(c_i|y_{<t}, X) - \beta H(p(c_i|y_{<t}, X)), \quad (11)$$

where β controls the strength of the confidence penalty, and $H(\cdot)$ is the entropy of the output distribution:

$$H(p(c_i|y_{<t}, X)) = -\sum_{i=1}^{N} p(c_i|y_{<t}, X) \log(p(c_i|y_{<t}, X)). \quad (12)$$

The authors also show that this confidence penalty method is closely related to label smoothing regularization (Szegedy et al., 2016), therefore methods like neighborhood smoothing (Chorowski and Jaitly, 2016) may be used to solve the low-diversity problem.

So far, there has been no published work on analyzing the the effectiveness of correcting for model over-confidence on the low-diversity problem. It is important to note the fourth diagnosis of the low-diversity problem, i.e., that the problem is due to model over-confidence, is essentially different from the three types of diagnosis that we described earlier in the section. Among diagnoses and methods published previously, the VAE-based approaches actually bypass the low-diversity problem by introducing randomness; MMI-based methods have an elegant theoretical basis, yet they end up relying on many extra modules, like reverse models and beam search, and the newly-introduced hyper-parameters were not even learned from training data (Li et al., 2015); attention-based models offer a complementary approach, since strengthening the conditional signal is likely to make the response more specific, which should in turn improve the corpus-level diversity. Model over-confidence may offer a simpler alternative – we believe that methods such as confidence penalty are likely to alleviate the low-diversity problem in ways that differ from previous approaches.

4 Next Steps

In this paper, we described the low-diversity problem for response generation, which is one of the main issues faced by current Seq2Seq-based conversation models. We reviewed existing diagnoses and corresponding approaches to this problem and also added a diagnosis that has not been proposed or used so far, i.e., model over-confidence.

By using entropy maximizing approaches, such as confidence penalty (Pereyra et al., 2017) or label smoothing (Szegedy et al., 2016), we believe that the low-diversity problem of Seq2Seq models can be alleviated. Besides, by using entropy maximizing methods, the self-repeating problem

(Li et al., 2017) of Seq2Seq models may also be alleviated since this can reduce the snowball effect and make later outputs more relevant. We also noticed that the low-diversity problem resembles the mode collapse problem of GANs (Goodfellow et al., 2014), therefore inspirations may be drawn from the solutions like (Salimans et al., 2016; Metz et al., 2016).

In addition, since we now have four types of diagnosis of the low-diversity problem, each of which is likely to address part of the problem but not all of the problem, it is natural to systematically compare and combine approaches based on the different types of diagnosis. Understanding how solutions to the low-diversity problem helps to improve the effectiveness of conversational agents for search-oriented tasks is another interesting line of future work.

Acknowledgments

This research was supported by the China Scholarship Council.

References

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

Jan Chorowski and Navdeep Jaitly. 2016. Towards better decoding and language model integration in sequence to sequence models. *arXiv preprint arXiv:1612.02695*.

Denis Fedorenko, Nikita Smetanin, and Artem Rodichev. 2017. Avoiding echo-responses in a retrieval-based conversation system. *arXiv preprint arXiv:1712.05626*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*.

Jiwei Li, Will Monroe, Tianlin Shi, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. 2016. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*.

Alexander H Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. 2017. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, volume 16, pages 3776–3784.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301.

Louis Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. *arXiv preprint arXiv:1701.03185*.

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.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. 2018. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In *IJCAI*, pages 4418–4424.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Oriol Vinyals, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015a. Grammar as a foreign language. In *Advances in Neural Information Processing Systems*, pages 2773–2781.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015b. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Jun Yin, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, and Xiaoming Li. 2015. Neural generative question answering. *arXiv preprint arXiv:1512.01337*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*.