Why are Sequence-to-Sequence Models So Dull?

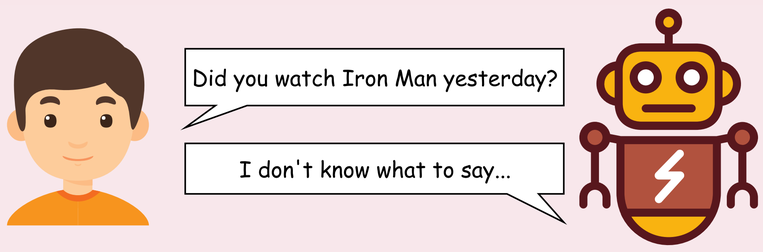
Understanding the Low-Diversity Problem of Chatbots

Shaojie Jiang Maarten de Rijke

University of Amsterdam

Motivation

Diversity is a long-studied topic in information retrieval. Sequence-to-sequence (Seq2Seq) based chatbots have similar problems: for many different user inputs, they just reply with generic responses, like "I don't know", "I'm sorry".



For this which we refer to as low-diversity problem, we care about:

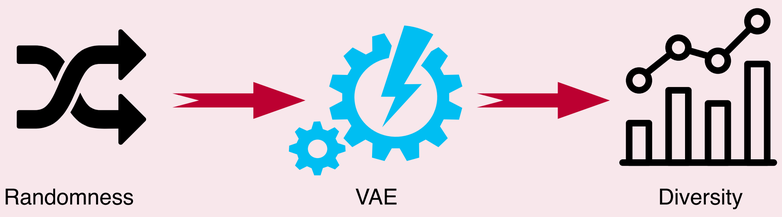
- What diagnoses and solutions have been proposed?
- Are there any other causes or solutions?

Existing diagnoses and solutions

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.

Solution:



Both works propose to introduce variational autoencoders (VAEs) to Seq2Seq models. At test time, the stochastic latent variable z from a VAE is inputed to the decoder LSTM:

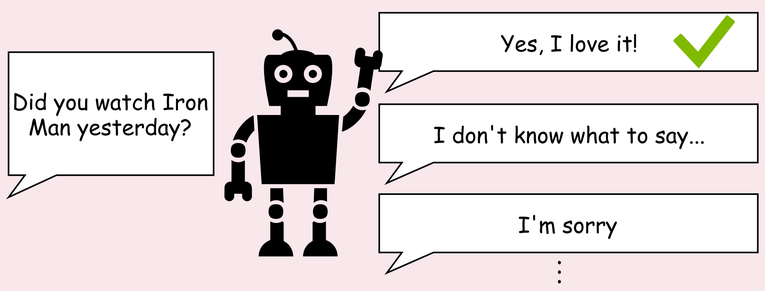
$$h_t^{dec} = f_\theta^{dec}(h_{t-1}^{dec}, y_{t-1}, z).$$

By this means, the variability brought by z is converted to response diversity. However, the underlying Seq2Seq model remains sub-optimal in terms of diversity.

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 $P(Y|X)$.

Solution:



The authors propose to maximize the mutual information (MMI) between input-output pairs (X, Y) by proposing two new objective functions:

$$\hat{Y} = \arg\max_Y \{\log P(Y|X) - \lambda \log P(Y) + \gamma|Y|\},$$

and

$$\hat{Y} = \arg\max_Y \{(1-\lambda)\log P(Y|X) + \lambda \log P(X|Y) + \gamma|Y|\}.$$

The theory is good, but in practice this method relies too much on beam search and a reverse model trained using (Y, X).

Weak conditional signal:

Tao et al. (2018) argue that the original attention signal often focuses on particular part of the input sequence, which is not strong enough for the Seq2Seq model to generate specific responses.

Solution:

They 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:

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

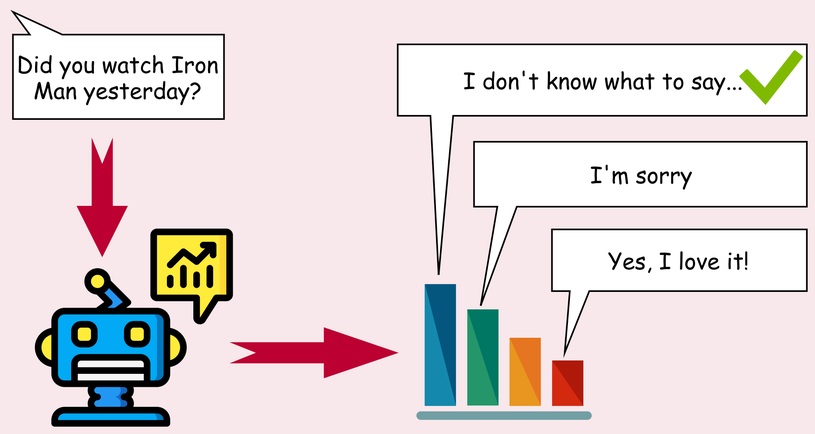
where $W_p^k \in \mathbb{R}^{d \times d}$ is a learnable projection matrix.

Our Diagnose

We found that the low-diversity problem of Seq2Seq-based chatbots has very close relationships with model over-confidence (Pereyra et al., 2017). Why?

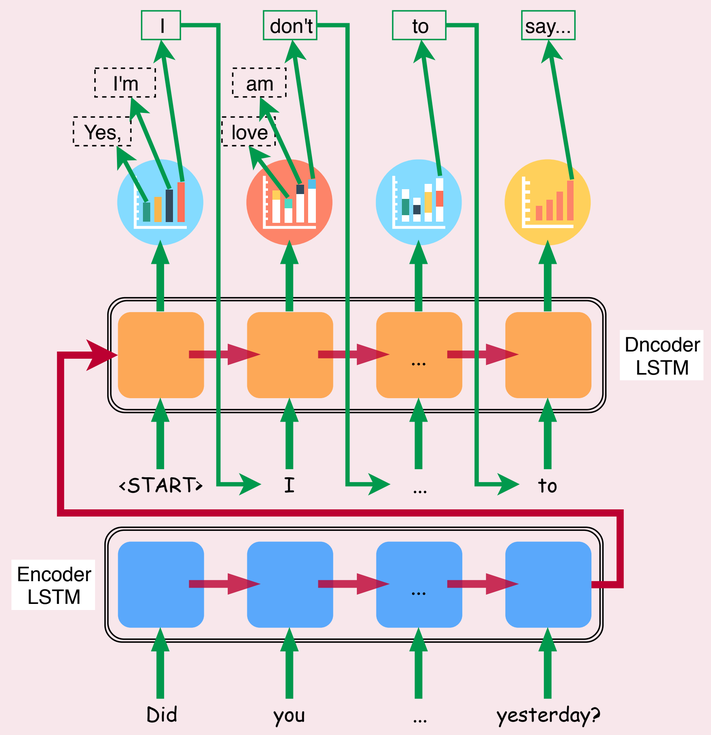
Nature of Seq2Seq-based chatbots

With the commonly used MAP objective function, the response with highest predicted probability is chosen as response:



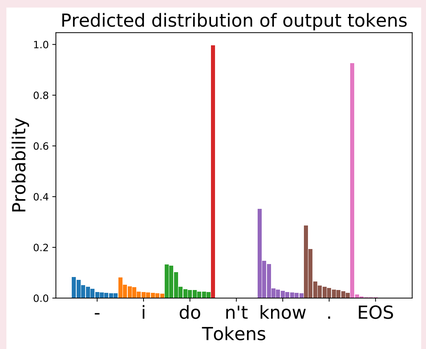
Details

Here's what happened inside the Seq2Seq chatbot:



An actual example:

By looking into an actual example, it is obvious that the model is becoming over-confident as the generation goes on:



Possible solution:

Although the reason of model over-confidence remains unclear, the approaches to this problem, such as confidence penalty and label smoothing (Pereyra et al., 2017), could be used to address the low-diversity problem.

Below is how we tailor the entropy maximizing loss function for Seq2Seq chatbots:

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

Conclusion

- We reviewed existing diagnoses and corresponding approaches to low-diversity problem of Seq2Seq-based chatbots
- We provided a different diagnose and pointed out possible solutions
- Performance of entropy maximizing loss function needs to be tested
- Reasons of model over-confidence need to be explored

UNIVERSITY OF AMSTERDAM

{s.jiang, derijke}@uva.nl