Neural Response Generation via GAN with an Approximate Embedding Layer

Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun,
Xiaolong Wang, Zhuoran Wang, and Chao Qi

Empirical Methods in Natural Language Processing, 2017

Problems Addressed

1. Applying GANs to NLP is problematic because GANs are only defined for real-valued data.
2. GANs work by training a generator network that outputs synthetic data, then running a discriminator network on the synthetic data. The gradient of the output of the discriminator network with respect to the synthetic data tells you how to slightly change the synthetic data to make it more realistic.
3. Slight changes can be made to the synthetic data only if it is based on continuous numbers. If it is based on discrete numbers, there is no way to make a slight change.
4. This problem can be tackled by using policy gradient, as is done in the case of SeqGANs by using REINFORCE algorithm. However, REINFORCE doesn't work very well in practice.
5. The proposed model forgoes the need to use policy gradient by introducing an Approximate Embedding Layer, which solves the non-differentiable problem caused by the sampling-based output decoding procedure in the Seq2Seq generative model.
6. The GAN setup provides an effective way to avoid non-informative responses (a.k.a "safe responses"), which are frequently observed in traditional neural response generators.

Proposed Solution

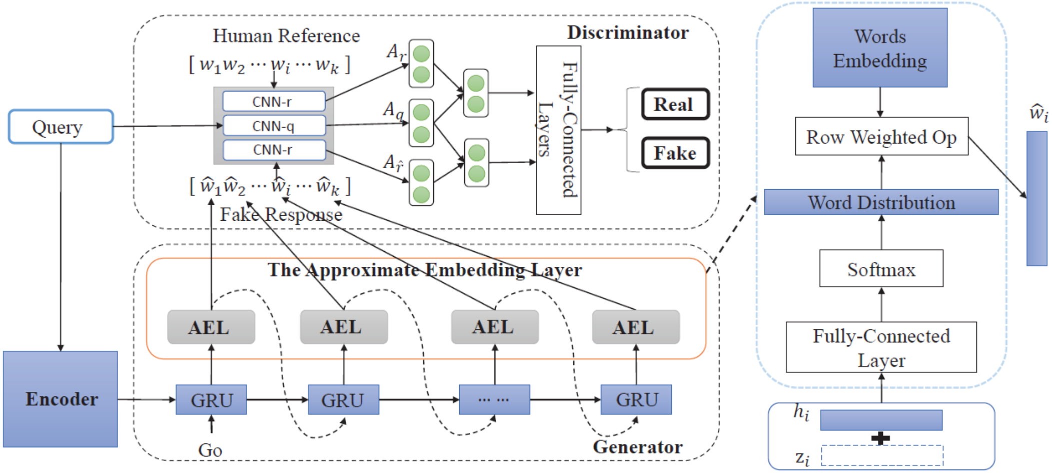

Figure 1: The Framework of GAN for the Response Generator.

Let

$$\mathbb{D} = \{(q_i, r_i)\}_{i=1}^N$$

be a set of N single turn human-human conversations, where

$$q_i = (w_{q_i,1}, ..., w_{q_i,t}, ..., w_{q_i,m})$$

is a query,

$$r_i = (w_{r_i,1}, ..., w_{r_i,t}, ..., w_{r_i,n})$$

stands for the response to q_i, and $w_{q_i,t}$ and $w_{r_i,t}$ denote the t^{th} words in q_i and r_i, respectively.

Generator

- The generator adopts the Gated Recurrent Unit (GRU) based encoder-decoder architecture.
- The generator is pre-trained by MLE, with the objective as:

$$\frac{1}{|\mathbb{D}|} \sum_{(q,r)\in\mathbb{D}} \sum_{t=1}^{K} \log p(w_{r,t}|q_v, w_{r,1}, ..., w_{r,t-1})$$

  where q_v is the encoded representation of query, as given by encoder.

  This guarantees that the generator produces grammatical utterances.

Approximate Embedding Layer

- An approximate embedding layer serves as an interface for the discriminator to propagate its loss to the generator.
- Instead of choosing the word vector corresponding to the word with the maximum softmax probability, the word probabilities in the distributions obtained from the decoder's Softmax layer are multiplied to the corresponding word vectors, to directly yield an approximately vectorized representation of the generated word sequences for use in the discriminator.
- This approximation is based on the assumption that ideally the word distributions should be trained to reasonably approach the one-hot representations of the discrete words.
- The approximation layer takes the output hi of the generator and a random noise z_i as the input, and reuses the word projection layer (pre-rained in the standard generator) to estimate the probability distribution of w_i.
- The noise z_i added to h_i forms a latent feature to enforce the diversity of the generated responses.
- The overall word embedding is calculated as:

$$\hat{e}_{w_i} = \sum_{j=1}^{V} e_j \cdot Softmax(W_p(h_i + z_i) + b_p)_j$$

  where W_p and b_p are the weight and bias parameters of the word projection layer, respectively.

Discriminator

- The CNN based discriminator is attached on top of the approximation layer, which aims to distinguish the fake responses output by the approximation layer and the corresponding human-generated references, conditioned on the input query.
- Let $\hat{r}$ stand for a sequence of word distributions projected from the hidden layers of the decoder RNN, based on which one would sample the output response utterance in a traditional Seq2Seq generator.

- The input of the discriminator consists of the following vectors, all zero padded or truncated to same length:

  a.) The word embedding vector sequence V_q for a given query q

  b.) The word embedding vector sequence V_r for its human-produced response r.

  c.) The approximate word embedding vector sequence $V_{\hat{r}}$ produced by the approximate embedding layer for the corresponding fake response $\hat{r}$.

- Two CNNs with shared parameters are employed to encode V_r and $V_{\hat{r}}$ into higher-level abstractions, termed A_r and $A_{\hat{r}}$.
- A separate CNN is used to abstract V_q as A_q.
- A_q is concatenated to A_r and $A_{\hat{r}}$ separately, and the two vectors are fed to two FC layers with shared parameters.
- The two FC layers predict probabilities $D(r|q)$ and $D(\hat{r}|q)$, respectively, for r and $\hat{r}$ being true responses of the given q.
- The Discriminator is pre-trained to maximise the following function, with the parameters of G frozen:

$$D_{loss} = \log D(r|q) + log(1 - D(\hat{r}|q))$$

Adversarial Training

- When training G, the objective function of Discriminator is replaced with the l_2-loss between A_r and $A_{\hat{r}}$, to maintain a reasonable scale of the gradient.
- Only the parameters of decoder's hidden layers are tuned; the parameters of the encoder network and the projection layer of the decoder network are frozen.

  This is based on the assumption that, in principle, after the pre-training, the encoder network is sufficiently effective to represent the entire input utterance, while the projection layer of the decoder is also adequate to decode words from its hidden states.

- Hence, Adversarial Training is to adjust the way G organises the semantic contents during the decoding (or in other words, the way it realises the hidden states).

Key Points

- Both the generator and the discriminator are conditioned on the input query, which guarantees the relevance of the generated responses.
- The discriminator forces the generator to produce a response according to the true distribution in better granularity, such that the state of promoting safe responses is leaped out.
- The approximation layer yields a smooth connection between the generator and the discriminator, avoiding the non-differentiable discrete sampling process.

Experiments

1. The model is tested on Baidu Tieba and OpenSubtitles datasets.

Results

Evaluation Metrics

- *Relevance Metrics*: The following three word embedding based metrics are used to compute the semantic relevance of two utterances.

  a.) *The Greedy metric* is to greedily match words in two given utterances based on the cosine similarities of their embeddings, and to average the obtained scores.

  b.) *The Average metric*, in which an utterance representation is obtained by averaging the embeddings of all the words in that utterance, and then the cosine similarity is computed.

  c.) *Extreme metric*, in which an utterance representation is obtained by taking the largest extreme values among the embedding vectors of all the words it contains, before computing the cosine similarities between utterance vectors.

- *Diversity Metrics*: To measure the informativeness and diversity of the generated responses, dist-1 and dist-2 metrics are followed, and a Novelty metric is introduced.

  a.) The dist-1 (dist-2) is defined as the number of unique unigrams (bigrams for dist-2). A common drawback of dist-1 and dist-2 is that in the computation, less informative words (such as "I", "is", etc.) are considered equally with those more informative ones.

  b.) *Novelty metric* is the number of infrequent words observed in the generated responses. All the words except the top 2000 most frequent ones in the vocabulary are considered as infrequent words.

  *.) The dist-1 and Novelty values are normalised by utterance length, and dist-2 is normalised by the total number of bigrams in the generated response.

Model	Relevance			Diversity		
	Average	**Greedy**	**Extreme**	**Dist-1**	**Dist-2**	**Novelty**
Seq2Seq	0.720	0.614	0.571	0.0037	0.0121	0.0102
MMI-anti	0.713	0.592	0.552	0.0127	0.0495	0.0250
Adver-REGS	0.722	0.660	0.574	0.0153	0.0658	0.0392
GAN-AEL	**0.736**	**0.689**	**0.580**	**0.0214**	**0.0963**	**0.0635**

Table 1: Relevance and diversity evaluation on the Tieba dataset.

Model	Relevance			Diversity		
	Average	**Greedy**	**Extreme**	**Dist-1**	**Dist-2**	**Novelty**
Seq2Seq	0.719	0.578	0.505	0.0054	0.0141	0.0045
MMI-anti	0.710	0.569	0.499	0.0175	0.0586	0.0097
Adver-REGS	0.726	0.590	0.507	0.0223	0.0725	0.0147
GAN-AEL	**0.734**	**0.621**	**0.514**	**0.0296**	**0.0955**	**0.0216**

Table 2: Relevance and diversity evaluation on the OpenSubtitles dataset.

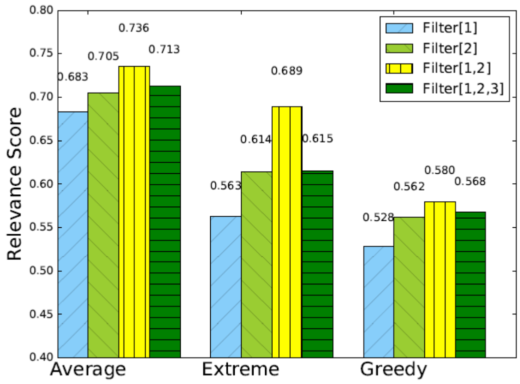

Figure 2: Relevance scores of GAN-AEL on the Tieba corpus with respect to different CNN window sizes.

GAN-AEL vs Adver-REGS		
Wins	Losses	Ties
0.61	0.13	0.26

Table 3: Evaluations of GAN-AEL and Adver-REGS based on human subjects,

Bibliography

1. Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In Advances in Neural Information Processing Systems 29, pages 2172–2180.

2. Xing Chen, Wu Wei, Wu Yu, Liu Jie, Huang Yalou, Zhou Ming, and Ma Wei-Ying. 2017. Topic aware neural response generation. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pages 3351–3357.

3. Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724–1734.

4. Ishan Durugkar, Ian Gemp, and Sridhar Mahadevan. 2016. Generative multi-adversarial networks. Proceedings of the 4th International Conference on Learning Representations (ICLR).

5. Gabriel Forgues, Joelle Pineau, Jean-Marie Larcheveque, and Real Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In NIPS, Modern Machine Learning and Natural Language Processing Workshop.

6. 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 27, pages 2672–2680.

7. Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In Advances in Neural Information Processing Systems 27, pages 2042–2050.

8. Serban Iulian, Vlad, Klinger Tim, Tesauro Gerald, Talamadupula Kartik, Zhou Bowen, Bengio Yoshua, and C. Courville Aaron. 2017. Multiresolution recurrent neural networks: An application to dialogue response generation. In Proceedings of the Thirty- First AAAI Conference on Artificial Intelligence, pages 3288–3294.

9. Yoon Kim. 2014. Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746–1751.

10. Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pages 110–119.

11. Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and William B. Dolan. 2016b. A persona-based neural conversation model. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL).

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

13. Pierre Lison and J¨org Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC).

14. Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2122–2132.

15. Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In Proceedings of ACL-08: HLT, pages 236–244.

16. Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In Proceedings of the 2011 Conference on Empirical Methods on Natural Language Processing (EMNLP), pages 583–593.

17. Vasile Rus and Mihai Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pages 157– 162.

18. Iulian Vlad Serban, Sordoni Alessandro, Lowe Ryan, Charlin Laurent, Pineau Joelle, C. Courville Aaron, and Bengio Yoshua. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pages 3295– 3301.

19. Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP), pages 1577– 1586.

20. 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. In Proceedings of the 14th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACLHLT), pages 196–205.

21. Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems 27, pages 3104–3112.

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

23. Bowen Wu, Baoxun Wang, and Hui Xue. 2016. Ranking responses oriented to conversational relevance in chat-bots. In Proceedings of the 26th International Conference on Computational Linguistics (COLING), pages 652–662.

24. Lantao Yu,Weinan Zhang, and and Yong Yu JunWang. 2017. SeqGAN: Sequence generative adversarial nets with policy gradient. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, volume 31.

25. Ganbin Zhou, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He. 2017. Mechanism-aware neural machine for dialogue response generation. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pages 3400–3407.

26. Xiangyang Zhou, Daxiang Dong, HuaWu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. 2016. Multi-view response selection for humancomputer conversation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 372–381.