Journal of Educational Media & Library Sciences 57 : 3 (2020) : 355-378
DOI:10.6120/JoEMLS.202011_57(3).0048.RS.CM

Research Article

Development and Evaluation of Emotional Conversation System Based on Automated Text Generation[ψ]

Te-Lun Yang[a] Yuen-Hsien Tseng[bc*]

Abstract

Based on the corpus provided by the 2019 Chinese Emotional Conversation Generation (CECG) evaluation task, an emotional conversation system is implemented in this paper using deep learning and other technologies such as GPT-2 and BERT. The effectiveness of the system is evaluated based on the test data and criteria provided by CECG. The results based on three human annotators show that the system has a similar effectiveness level with that of the best team participating in the 2019 CECG task. Further case studies reveal that the more post/reply pairs about a topic in the training data, the better the language model of GPT-2 to generate innovative, interesting, and perfect response sentences for that topic. The main contributions of this study are: 1. Integrating emotion into the post string as a condition for computing probability, so as to simply train GPT-2 and make GPT-2 predict in the original way; 2. Applying BERT to predict the coherence of response sentences as a basis for ranking. Although these two techniques are derived from the training mechanisms of GPT and BERT respectively, we have slightly modified them to fit the task of CECG and achieved good results.

Keywords: *Conversational system, Text generation, Text understanding, Deep learning, Artificial intelligence*

SUMMARY

Introduction

In human-computer interaction, automatic recognition of human emotions for appropriate response can make human-computer interaction smoother and more effective. Related research shows that the expression of empathy can increase user satisfaction and promote positive interaction.

[ψ] Both authors have the same contribution. Te-Lun Yang implemented the whole system, while Yuen-Hsien Tseng proposed the solution and complete the paper writing.

[a] Master Student, Graduate Institute of Library and Information Studies, National Taiwan Normal University, Taipei, Taiwan

[b] Distinguished Professor and Associate Dean, Graduate Institute of Library and Information Studies, National Taiwan Normal University, Taipei, Taiwan

[c] Co-Principal Investigator, MOST AI Biomedical Research Center

[*] To whom all correspondence should be addressed. E-mail: samtseng@ntnu.edu.tw

In pursuit of the above delicate interaction, this paper presents the building of a Chinese dialogue system that emphasizes on emotional conversation using state-of-the-art AI techniques, namely Generative Pre-trained Transformer (GPT) and Bidirectional Encoder Representations from Transformers (BERT). This emotional conversation system (ECS) is expected to respond to a user's post with a fluent and coherent reply conforming to a specified or detected emotion.

Problem

Specifically, this paper adopts the datasets and evaluation criteria from the Chinese Emotional Conversation Generation (CECG) Shared Task held in Short Text Conversation Task (STC-3) in the 14th NTCIR Workshop (2018-2019) to train the proposed system and evaluate its performance.

The CECG task is defined as: for a user input post, the system needs to output a response (or reply) with a specified emotion category, of which there are 5 types of emotion: Anger, Disgust, Happiness, Like, and Sadness. A possible example of the post/reply is as follows:

User's post: My cat died yesterday.

System reply (given a specified emotion type in the squared bracket):

[1: like] Oh she likes to make believe. That's cute.

[2: Sadness] Oh, I'm so sorry for your loss.

[3: Disgust] That's fine. That would save you a lot of trouble.

[4: Anger] Was it killed? Let's find out who did it!

[5: Happiness] How fortunate! She's an angel now in heaven.

Note that in this imaginative example, the user's post may express sadness, it is rather difficult to make a like response, even for human.

The datasets provided by the CECG Shared Task are based on the pairs of posts and responses of Weibo users mainly from mainland China, with a total of about 1.7 millions of pairs (about 1.1 millions of pairs in 2019 and about 600,000 in 2017). For each post or replied text, a machine classifier was trained to label the emotion type of the text, and its accuracy is about 62%.

The CECG Shared Task evaluates each reply based on the post and the specified emotion, by human, according to the following criteria:

```
IF Coherence and Fluency
    IF Emotion Consistency
        LABEL 2
    ELSE
        LABEL 1
ELSE
        LABEL 0
```

Note that Coherence means that the reply is consistent with the topic of the post, Fluency means that the reply text is smooth and grammatically correct, and Emotion Consistency denotes that the reply's emotion is consistent with the specified emotion.

Method

The developed ECS system consists of a user interface, a GPT-2 model for text generation, and a BERT model for text understanding and coherence prediction. The ECS takes input post from users through a Web API (or Web UI). An open source GPT-2 Chinese model was trained to output k candidate replies. These candidates were further ranked by a Chinese BERT model trained to predict the coherence of the reply based on the post. The highest ranked candidate was then chosen as the output reply.

The training data from CECG were in the form: [[post_i, post_i_emotion], [reply_i, reply_i_emotion]]. They were converted into the form: ["post_i [reply_i_emotion] reply_i"] so as to conform to the training data format of GPT-2. In other word, the CECG problem asks us to predict the reply based on two conditions:

$$P(\text{ reply} \mid \text{post, emotion})$$

By concatenating the two conditions into one string, we reformulated the problem into the original language model learnable and predictable by GPT-2:

$$P(\text{ reply} \mid \text{"post [emotion]"})$$

The GPT-2 was trained on a Titan RTX GPU with 24GB RAM. It took approximately 200 hours to train the 1.7 millions of post/reply pairs for 100 epochs.

The GPT-2 can be configured to output k candidate replies. To rank these candidates, a Chinese BERT pretrained model from Google was downloaded and fine-tuned on part of the original training data with the following format for coherence prediction:

```
[
  [ post_1, reply_1, 1.0 ],
  [ post_3, reply_7, 0.0 ],
  [ post_8, reply_8, 1.0 ],
  [ post_9, reply_2, 0.0 ],
…
]
```

In other words, the BERT model was trained to do linear regression prediction: if the input is the original post/reply pairs from the training data, the desired output has a score of 1.0; if the input is the scrambled post/reply pairs, the desired output is 0.0 to indicate that the post and reply are not coherent. The BERT model was fine-tuned on 15,000 pairs for one epoch (about 3 minutes), in which paired and scrambled post/reply are 50% each.

Findings

Based on the evaluation criteria of the CECG Shared Task in 2019, the evaluation of 1,000 ECS generated replies by three native speakers majored in Chinese linguistics indicated that 90.3% reply texts are grammatical correct, and 59.1% are coherent to the posted text, and about 88% reply sentences are novel (not in the 1.7M training texts). This result outperformed the top-ranking system in the 2019 task, where a hybrid method of using both text generation and rule-based mechanisms was applied. Further case studies revealed that the ECS could generate innovative, interesting, and perfect response sentences for popular topics in the training data.

As an example, for the post stated "I would like to be with you forever," example replies would look like: "I would also like to be with you forever." if the specified emotion is "Like"; and "Why do you have to stay with me? It's not fair!" if the specified emotion is "Disgust".

More exploration of the ECS showed that, If the topic of the post is rich in the training data, the GPT-2 can generate creative sentences; if the topic of the post is relatively scarce in the training data, the smoothness and topic coherence of the generated sentence will diminish. These results are similar with conclusions from previous studies.

Conclusions

The main contributions of this study are: 1. Integrating emotion type into the post text as a single condition for language modeling, so as to train and apply GPT-2 in the original way; 2. Applying BERT to predict the coherence of response text for ranking the generated replies. Although these two techniques are derived from the training mechanisms of GPT and BERT, respectively, we have slightly modified the techniques to fit the task of CECG and achieved good results.

This work sheds light on the pursuit of delicate human-computer interaction with emotion. Future work is needed to achieve the goal of better response texts (through better language modeling or larger training dataset) and to propose effective response strategies to yield proper emotional reply once a corresponding emotion was detected in the post.

ROMANIZED & TRANSLATED REFERENCE FOR ORIGINAL TEXT

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*, 1137-1155.

Binsted, K. (1995). *Using humour to make natural language interfaces more friendly* [Paper presentation]. Workshop on AI, ALife and Entertainment, International Joint Conference on Artificial Intelligence, Montreal, Canada.

Binsted, K., Bergen, B., Coulson, S., Nijholt, A., Stock, O., Strapparava, C., Ritchie, G., Manurung, R., Pain, H., Waller, A., & O'Mara, D. (2006). Computational humor. *IEEE Intelligent Systems*, *21*(2), 59-69. https://doi.org/10.1109/MIS.2006.22

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., …, Amodei, D. (2020). *Language models are few-shot learners*. arXiv. https://arxiv.org/abs/2005.14165v4

Cagan, T., Frank, S. L., & Tsarfaty, R. (2017). Data-driven broad-coverage grammars for opinionated natural language generation (ONLG). In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th annual meeting of the Association for Computational Linguistics* (Vol. 1, pp. 1331-1341). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1122

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv. https://arxiv.org/abs/1810.04805v1

Du, Z. (2019). GPT2-Chinese: Tools for training GPT2 model in Chinese language. Retrieved January 12, 2020, from https://github.com/Morizeyao/GPT2-Chinese

Ghosh, S., Chollet, M., Laksana, E., Morency, L.-P., & Scherer, S. (2017). Affect-LM: A neural language model for customizable affective text generation. In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th annual meeting of the Association for Computational Linguistics* (Vol. 1, pp. 634-642). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1059

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251-257. https://doi.org/10.1016/0893-6080(91)90009-T

Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., & Xing, E. P. (2017). Toward controlled generation of text. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (PMLR 70, pp. 1587-1596). ML Research Press.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Proceedings of the 25th International conference on neural information processing systems* (pp. 1097-1105). Neural Information Processing Systems Foundation.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning*. *Nature*, *521*, 436-444. https://doi.org/10.1038/nature14539

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L.

Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Proceedings of the 26th international conference on neural information processing systems* (Vol. 2, pp. 3111-3119). Neural Information Processing Systems Foundation.

Minsky, M., & Papert, S. A. (1969). *Perceptrons: An introduction to computational geometry*. MIT Press.

Partala, T., & Surakka, V. (2004). The effects of affective interventions in human-computer interaction. *Interacting with Computers*, *16*(2), 295-309. https://doi.org/10.1016/j.intcom.2003.12.001

Prendinger, H., & Ishizuka, M. (2005). The empathic companion: A character-based interface that addresses users' affective states. *Applied Artificial Intelligence: An International Journal*, *19*(3-4), 267-285. https://doi.org/10.1080/08839510590910174

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations*. MIT Press.

Skowron, M. (2010). Affect listeners: Acquisition of affective states by means of conversational systems. In A. Esposito, N. Campbell, C. Vogel, A. Hussain, & A. Nijholt (Eds.), *Lecture notes in computer science: Vol. 5967. Development of multimodal interfaces: Active listening and synchrony* (pp. 19-181). Springer. https://doi.org/10.1007/978-3-642-12397-9_14

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Proceedings of the 27th international conference on neural information processing systems* (Vol. 2, pp. 3104-3112). Neural Information Processing Systems Foundation.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Proceedings of the 30th international conference on neural information processing systems* (pp. 6000-6010). Neural Information Processing Systems Foundation.

Zhang, Y., & Huang, M. (2019). *Overview of the NTCIR-14 short text generation subtask: Emotion generation challenge* [Paper presentation]. 14th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo Japan.

Zhou, H., Huang, M., Zhang, T., Zhu, X., & Liu, B. (2018). Emotional chatting machine: Emotional conversation generation with internal and external memory. In *The thirty-second AAAI conference on artificial intelligence* (pp. 730-738). Association for the Advancement of Artificial Intelligence.

Te-Lun Yang ORCID 0000-0002-3351-1785
Yuen-Hsien Tseng ORCID 0000-0001-8904-7902