Response Selection for a Virtual Counsellor

Lee, John; Liang, Baikun

Published in:
The Web Conference 2021

Published: 01/04/2021

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

License:
CC BY

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.1145/3442442.3452312

Publication details:
Lee, J., & Liang, B. (2021). Response Selection for a Virtual Counsellor. In *The Web Conference 2021: Companion of The World Wide Web Conference WWW 2021* (pp. 495-499). (The Web Conference - Companion of the World Wide Web Conference, WWW). Association for Computing Machinery, Inc.
https://doi.org/10.1145/3442442.3452312

Response Selection for a Virtual Counsellor

John Lee, Baikun Liang
Department of Linguistics and Translation
City University of Hong Kong
Hong Kong SAR, China
jsylee@cityu.edu.hk,baikliang2-c@my.cityu.edu.hk

ABSTRACT

Chatbots are increasingly used for delivering mental health assistance. As part of our effort to develop a chatbot on academic and social issues for Cantonese-speaking students, we have constructed a dataset of 1,028 post-reply pairs on test anxiety and loneliness. The posts, harvested from Cantonese social media, are manually classified to a symptom category drawn from counselling literature; the replies are human-crafted, offering brief advice for each post. For response selection, the chatbot predicts the quality of a candidate post-reply pair with a regression model. During training, the symptom categories were used as proxies of reply relevance. In experiments, this approach improved response selection accuracy over a binary classification model and a weakly supervised regression model. This result suggests that manual annotation of symptom category can help boost the performance of a counsellor chatbot.

CCS CONCEPTS

• **Information systems → Language models**; **Retrieval models and ranking**.

KEYWORDS

chatbots, response retrieval, counselling, anxiety, loneliness, Cantonese

ACM Reference Format:
John Lee, Baikun Liang. 2021. Response Selection for a Virtual Counsellor. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3442442.3452312

1 INTRODUCTION

As virtual agents for mental health become evermore widespread, a number of counsellor chatbots have already been deployed [4, 9, 13]. While chatbots may not qualify to replace human counsellors, users can be more willing to disclose information to a chatbot than to a human [15]. The amount of publicly available counselling conversation data remains limited, however, for training chatbots to address specific counselling topics in low-resource languages.

With the goal of developing a chatbot on academic and social issues for Cantonese-speaking students, we have constructed a dataset of 1,028 post-reply pairs on test anxiety and loneliness.[1] The posts were retrieved from Cantonese social media and tagged with a counselling issue (e.g., "I have nobody to talk to"), which we will refer to as a "symptom category", drawn from counselling research literature (Table 1). The replies were manually composed to offer brief advice for each post. This paper introduces this dataset and investigates the utility of the symptom categories for the *response selection* task: given a user post and a set of candidate replies, select the reply with optimal relevance to the post.

A common approach for response selection is to train a binary classifier, with gold replies as positive examples and randomly selected replies as negative examples [19, 23]. We adopt a regression model to predict reply relevance, and train the model with symptom categories as proxies for the degree of matching between post and reply. Experiments show that our proposed approach achieves higher accuracy in response selection over a binary classifier and a weakly supervised regression model. This result suggests that manual labeling of symptom category can serve as a lower-cost annotation to boost the performance of a counsellor chatbot.

The rest of the paper is organized as follows. After a summary of previous work (Section 2), we present our dataset (Section 3). We then describe our approach (Section 4) and report experimental results (Sections 5 and 6).

2 PREVIOUS WORK

We summarize previous research in counsellor chatbot developments (Section 2.1) and in response selection in chatbots in general (Section 2.2).

2.1 Counsellor chatbot development

While transcripts of counselling conversations can be found in published books, there are few publicly available electronic resources for training data-driven chatbots [6, 10]. For the topic of test anxiety, a small-scale corpus has been constructed [12], but it consists of simulated posts rather than genuine ones from social media.

For the topic of loneliness, many counsellor chatbots make use of FAQ and knowledge bases. For example, PAL [13] matches user input on a number of social issues with information in the *Baidu Zhidao* knowledge base, and then selects the best answer matching the user's gender, marital status and age. Similarly, TeenChat [9] detects the type of user's stress, and replies with a sentence in the corresponding category in a knowledge base.

WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3452312

[1] Our dataset is publicly available at:
https://github.com/CantoneseCounsellorChatbot/Cantonese-Counsellor-Chatbot

Post	Reply	Topic	Symptom category
我唔開心，好似真係無人會理咁 I'm so sad, nobody cares about me	明白呢個感受會令人唔開心，試下對人敞開自己 I know this makes you unhappy; try open up to others	Loneliness	I have nobody to talk to
冇人聽我講野，乜都要自己面對 No one talks to me, I have to face everything by myself	你可以平時同身邊人建立良好關係，我相信佢地都會肯幫你同聽你講野 You can establish relationships with others around you, I think they'll listen to you	Loneliness	I have nobody to talk to
有左女朋友之後就無哂朋友 No friend after I started dating	或者試下多同同學開新話題 Maybe find new topics with other students	Loneliness	I lack companionship
我會答錯哂，然後我啊媽會鬧 I'll get everything wrong and my mom will scold me	不要讓父母成為了你的壓力,盡了力便可 Don't let your parents become your stress, just try your best	Test anxiety	I worry about what my parents will say

Table 1: Example post-reply pairs in our dataset, each labelled with its topic and symptom category, which is drawn from the lists in Tables 3 and 4

2.2 Response selection in chatbots

Most chatbots follow either the generation-based or retrieval-based paradigm to construct a reply for a user post. The former, based on natural language generation techniques, can create a response tailored to the post [21, 22, 27]. Possible drawbacks, however, include unpredictability of the generated reply, and inconsistencies in content or style [5]. These can lead to negative consequences for the counsellee due to an inappropriately generated advice.

Small talk is an essential component for rapport building in a counselling session [11]. A possible division of labor, then, is to perform small talk with the generation-based paradigm, while offering advice with the retrieval-based paradigm. The retrieval-based method, which selects the most relevant response from an existing database of candidates [1, 16], has been widely used in both task-oriented [18] and general-purpose chatbots [28].

A response selection algorithm may consider the semantic relatedness between the user input and existing posts in the training data [18], and/or the relevance of a candidate reply, in other words the degree of matching between the post and a candidate reply. This task may be viewed as binary classification, with human responses labelled as positive examples and randomly sampled ones as negative examples. BERT [3], for example, can be fine-tuned to perform the classification [19, 23]. A response selection model can also be pre-trained on a large general-domain conversational corpora, and then fine-tuned with a smaller set of in-domain post-reply pairs [8].

An alternative approach is to predict the matching score of post-reply pairs. To avoid manual scoring of a large number of post-reply pairs as training data, a sequence-to-sequence model trained on the general domain can serve as a weak annotator to estimate the scores of unlabelled pairs [26]. Experimental results indicate that this weakly supervised approach yields significant improvement over binary classification. Pursuing this line of research, our study investigates the use of symptom category as a proxy for the relevance of a candidate reply for a post.

3 DATASET

Cantonese is considered the "most widely known and influential variety of Chinese other than Mandarin" [17]. Less frequently used in formal written communication than Mandarin, the dominant variety in mainland China, Cantonese is supported by relatively few resources for natural language processing. The two are mutually unintelligible in their spoken form and have significant differences in their written form [24].

Topic	# post-reply pairs	# chars/post	# chars/reply
Test anxiety	570	17.2	26.9
Loneliness	458	35.0	39.0
Total	1,028	25.1	32.3

Table 2: Statistics on our dataset, with a breakdown into its two topics

We now describe the three stages in the construction of our dataset, which consists of 1,028 Cantonese post-reply pairs on two common counselling topics for university students: test anxiety and loneliness (Table 2).

Post collection. We recruited 10 undergraduate students to collect sentences from Cantonese social media, such as Facebook and Instagram, that discuss loneliness and academic anxiety in general. The students collected a total of 4,744 posts.

Post labeling. The students labelled each post with a counselling issue, which we refer to as a "symptom category", addressed in counselling literature. We used the 26 symptom categories identified by Wren and Benson [25] for the topic of test anxiety (Table 3), and the 20 symptoms identified by Russell et al. [20] for the topic of loneliness (Table 4). The symptom category of each post was independently labelled by two annotators. In case of disagreement, a third annotator performed the labeling. If all three labels differed, the sentence was excluded from the dataset. Otherwise, the sentence was assigned the symptom category with the majority vote.

Reply composition. Finally, the students composed a reply for each post, typically acknowledging its main point and offering brief advice. Staff at the counselling department at our university offered guidance with example replies.

4 APPROACH

We compare three methods for constructing training data for the *response selection* task: given a user post, the chatbot is to select the best reply from the dataset (Section 3). The training data consists of

Symptom group	# posts
I think I am going to get a bad grade I think about what my grade will be	69
I think about what will happen if I fail I worry about failing	46
I worry about what my parents will say	70
I worry about doing something wrong I think most of my answers are wrong I think about how poorly I am doing	49
I worry about how hard the test is I think that I should have studied more It is hard for me to remember the answers	94
I look around the room I look at other people I stare	23
I check the time I try to finish up fast	37
I tap my feet I play with my pencil	22
I feel nervous I feel scared	75
My heart beats fast My head hurts I feel warm My hand shakes My belly feels funny	74
I do not want to live any more	11

Table 3: The 26 symptom categories for the topic of test anxiety [25], manually clustered into 11 groups for regression model training (Section 4.1)

triplets $\{(p_i, r_i, y_i)\}$, where p_i is a user post; r_i is a candidate reply; and y_i is the relevance score that quantifies the degree of matching between p_i and r_i.

4.1 Symptom category-based regression

We trained a regression model to predict the relevance of a reply for a post. To avoid labor-intensive manual scoring of a large number of post-reply candidates as training data, we used symptom categories as proxies of the relevance score. Intuitively, reply relevance is often correlated with its symptom category. Consider the reply to the post in the first row in Table 1. It may be highly relevant to a post belonging to the same symptom category, such as the one in the second row ("I have nobody to talk to"); it is likely still somewhat relevant to a post from another category within the same topic, such as the one in the third row ("I lack companionship"); but it is unlikely to be suitable for a post in another topic (fourth row).

To create a more fine-grained hierarchy, we manually clustered the symptom categories into groups based on their relatedness. The 26 symptom categories for test anxiety were clustered into 11 symptom groups (Table 3), and the 20 symptom categories for loneliness into 8 groups (Table 4).

The assignment of relevance scores are shown in Table 5. The gold reply is assigned $y_i = 1$. To train the model to distinguish between semantically close replies, we included all non-gold replies in the same symptom category, with the assigned relevance score of $y_i = 0.75$; and all replies from the same symptom group, with $y = 0.50$.[2] Further, we randomly selected M replies from the same topic ($y_i = 0.25$), and M replies from a different topic ($y_i = 0$). We will report experimental results with the settings $M = \{5, 10, 20\}$.

Symptom group	# posts
I am unhappy doing so many things alone I cannot tolerate being so alone	46
I have nobody to talk to I find myself waiting for people to call or write There is no one I can turn to	57
I lack companionship I feel completely alone People are around me but not with me	96
I feel as if nobody really understands me No one really knows me well	37
My interests and ideas are not shared by those around me I am unable to reach out and communicate with those around me	74
I am no longer close to anyone My social relationships are superficial	42
I feel left out I feel isolated from others I feel shut out and excluded by others	35
I feel starved for company I am unhappy being so withdrawn It is difficult for me to make friends	71

Table 4: The 20 symptom categories for the topic of loneliness [20], manually clustered into 8 groups for regression model training (Section 4.1)

Reply type	Score
Gold reply	1
Candidate reply in same symptom category	0.75
Candidate reply in same symptom group	0.50
Candidate reply in same topic	0.25
Candidate reply in different topic	0

Table 5: Assignment of relevance score to candidate replies in the symptom category-based regression model (Section 4.1)

4.2 Weakly supervised regression

A regression model for predicting reply relevance can also be trained with the relevance scores predicted by a pre-trained sequence-to-sequence model. Adopting an approach similar to [26], we used a pre-trained response generation model in the open domain to predict the relevance score for the non-gold replies.

[2] Using only a subset of the non-gold replies from the same symptom category and symptom group resulted in worse performance.

Model	Parameter	p@1	p@5	p@10	R_{100}@1	R_{100}@5	R_{100}@10
Binary classification (Section 4.3)	$N = 3$	0.303	0.456	0.496	0.464	0.608	0.703
	$N = 4$	0.304	0.462	0.518	0.467	0.642	0.762
	$N = 5$	0.202	0.380	0.468	0.398	0.610	0.713
Weakly supervised regression (Section 4.2)	$M = 10$	0.192	0.428	0.519	0.457	0.667	0.738
Symptom category-based regression (Section 4.1)	$M = 5$	0.350	0.488	0.564	0.505	0.751	0.834
	$M = 10$	**0.370**	**0.518**	**0.585**	**0.529**	**0.755**	**0.865**
	$M = 20$	0.356	0.491	0.559	0.518	0.743	0.842
	gold symptom	0.387	0.558	0.651	0.584	0.771	0.887

Table 6: Performance of the three response selection models described in Section 4

Specifically, we made use of GPT2-chitchat[3], a GPT-2 model that has been pre-trained on over 0.5M Mandarin dialogs in a similar manner as DialoGPT [27]. The gold reply is assigned $y_i = 1$. For non-gold replies, we computed $loss(r_i)$, which is the cross-entropy loss between r_i and the reply generated by the model, and then assigned $y_i = \max(0, 1 - \frac{loss(r_i)}{loss(gold)})$, to normalize the bias from different p_i.

4.3 Binary classification

A common response selection approach is to train a binary classifier to label a reply as relevant or not relevant for a post [19, 23]. As another baseline, we followed this approach and assigned $y_i = 1$ when r_i is the gold reply to p_i. We then randomly selected N non-gold replies as negative examples and assigned $y_i = 0$. The optimal number of N may vary according to dataset, ranging from 4 to 20 [23, 26]. We tried the settings $N = \{1, 2, 3, 4, 5, 10\}$ and report results for the three best parameter values.

5 EXPERIMENTAL SET-UP

We used chinese-roberta-wwm-ext [2], which is pre-trained with whole-word masking in the RoBERTa framework [14], a state-of-the-art neural language model. We fine-tuned this model to perform regression (Sections 4.1 and 4.2)[4] and classification (Section 4.3)[5] using transformers.

We adopt the following evaluation metrics for response selection [7, 8, 29]:

- $p@K$ Proportion of posts for which the gold reply is among the top-ranked K out of all candidate replies in the dataset (Table 2).
- $R_{100}@K$ Proportion of posts for which the gold reply is the top-ranked K out of 100 candidate replies randomly selected from the dataset.

We performed further evaluation on the more realistic setting where the chatbot does not necessarily reply with an advice for each user input. When no relevant reply can be found with high confidence, for example, it may alternatively generate encouragers, such as restatement or short questions [11], to sustain the conversation. In this setting, precision is defined as the number of gold replies selected, divided by the number of attempted replies; recall is defined as the number of gold replies selected, divided by the number of posts in the test set. We report on $F_{0.5}$, which gives twice as much weight on precision over recall, to reflect the importance of high-quality responses in the counselling task to avoid adverse effects on the user.

6 RESULTS

Table 6 reports the performance of the three response selection models described in Section 4, based on five-fold cross validation on our dataset (Section 3).

Baselines. The binary classifier (Section 4.3) performed best when trained with $N = 4$, i.e., four negative replies per post, attaining $p@1$ at 0.304.

Weakly supervised regression (Section 4.2) achieved lower $p@1$ (0.192) and $R_{100}@1$ (0.457) than the binary classification model. It outperformed the classifier, however, in terms of $p@10$ and $R_{100}@10$. This suggests that an open-domain dialog model can be more effective in detecting generally relevant replies, but less so in determining the best one.

Proposed approach. The best result was obtained by the symptom category-based regression model (Section 4.1) at $M = 10$, i.e. trained with 10 negative samples from different symptom groups and different topics. This setting also produced the highest $p@K$ and $R_{100}@K$ for all K values. Its $p@1$, at 0.370, represents a substantial improvement over the binary classifier. This suggests that manual annotation of symptom category, while requires relatively less effort, can boost response selection accuracy.

Figure 1 shows the precision and recall at various confidence thresholds. The proposed model achieved an optimal $F_{0.5}$ of 0.756, at 0.790 precision and 0.644 recall.

Finally, we measured the ceiling performance of the symptom category-based regression model, when it has access to the gold symptom category. The $p@1$ of the model further increased to 0.387. This suggests that while accurate symptom classification can further improve performance, distinguishing the nuances between posts within the same symptom category remains a challenge.

[3] https://github.com/yangjianxin1/GPT2-chitchat

[4] Using *Simple Transformers* (https://simpletransformers.ai/), we fine-tuned the regression model in Section 4.1 for 7 epochs, as determined by the validation dataset. We used the Adam optimizer, a learning rate of 4e-5, a training batch size of 32, and mean square error as the loss function. The same implementation and settings were used for fine-tuning the regression model in Section 4.2.

[5] Using *Simple Transformers* (https://simpletransformers.ai/), we fine-tuned the sentence-pair classification model with the Adam optimizer, a learning rate of 4e-5, a training batch size of 32, with the accuracy score as loss function.

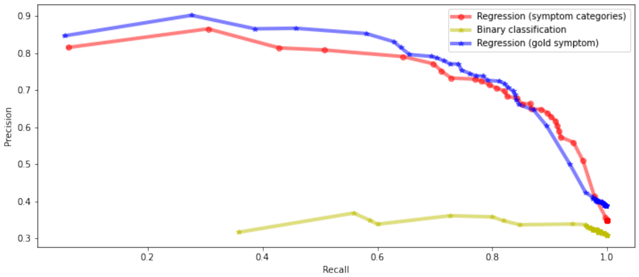

Figure 1: Precision and recall of the binary classification model (Section 4.3), the symptom category-based regression model (Section 4.1), and the version of this model with access to the gold symptom category

7 CONCLUSION

To meet the growing needs for mental health, chatbots can be expected to play larger roles in delivering counselling services. To support chatbot development, we have constructed a dataset of 1,028 Cantonese post-reply pairs, consisting of posts from social media that discuss test anxiety and loneliness, and human-crafted replies that offer brief advice. Each post is classified into a "symptom category", an issue drawn from the counselling literature.

We trained a regression model to predict the quality of post-reply pairs. We conducted experiments to investigate whether symptom categories can serve as proxies for the relevance of a reply to a post. This approach achieved more accurate response selection, outperforming a weakly supervised model and a binary classification model. This result suggests that symptom category annotation can boost the quality of a virtual counsellor.

ACKNOWLEDGMENTS

This work is supported by a grant from the Health and Medical Research Fund (project #17180961), the Food and Health Bureau, The Government of the Hong Kong Special Administrative Region.

REFERENCES

[1] A. Bartl and G. Spanakis. 2017. A Retrieval-Based Dialogue System Utilizing Utterance and Context Embeddings. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 1120–1125.

[2] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-Training with Whole Word Masking for Chinese BERT. In *arXiv:1906.08101*.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. In *Proc. NAACL-HLT*.

[4] K. K. Fitzpatrick, A. Darcy, and M. Vierhile. 2017. Delivering Cognitive Behavior Therapy to Young Adults with Symptoms of Depression and Anxiety using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health* 4, 2 (2017), e19.

[5] Xiang Gao, Yizhe Zhang, Sungjin Lee, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2019. Structuring Latent Spaces for Stylized Response Generation. In *Proc. EMNLP*.

[6] Chulaka Gunasekara, Jonathan K. Kummerfeld, Lazaros Polymenakos, and Walter Lasecki. 2019. DSTC7 task 1: Noetic end-to-end response selection. In *Proc. First Workshop on NLP for Conversational AI*.

[7] Matthew Henderson, Iñigo Casanueva, Nikola Mrksic, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulic. 2019. ConveRT: Efficient and Accurate Conversational Representations from Transformers. *CoRR* abs/1911.03688 (2019). arXiv:1911.03688 http://arxiv.org/abs/1911.03688

[8] Matthew Henderson, Ivan Vulić, Daniela Gerz, I nigo Casanueva, Pawel Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019. Training Neural Response Selection for Task-Oriented Dialogue Systems. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

[9] Jing Huang, Qi Li, Yuanyuan Xue, Taoran Cheng, Shuangqing Xu, Jia Jia, and Ling Feng. 2015. TeenChat: A Chatterbot System for Sensing and Releasing Adolescents' Stress. *LNCS* 9085 (2015), 133–145.

[10] Masashi Inoue, Ryoko Hanada, Nobuhiro Furuyama, Toshio Irino, Takako Ichinomiya, and Hiroyasu Massaki. 2012. Multimodal Corpus for Psychotherapeutic Situation. In *Proc. LREC Workshop on Multimodal Corpora for Machine Learning*. 18–21.

[11] Allen E. Ivey and Mary Bradford Ivey. 2003. *Intentional Interviewing and Counseling: Facilitating Client Development in a Multicultural Society.* Brooks Cole.

[12] John Lee, Tianyuan Cai, Wenxiu Xie, and Lam Xing. 2020. A Counselling Corpus in Cantonese. In *Proc. Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*.

[13] Yuanchao Liu, Ming Liu, Xiaolong Wang, Limin Wang, and Jingjing Li. 2013. PAL: A Chatterbot System for Answering Domain-specific Questions. In *Proc. 51st Annual Meeting of the Association for Computational Linguistics (ACL)*. 67–72.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[15] G. M. Lucas, J. Gratch, A. King, and L. P. Morency. 2014. It's Only a Computer: Virtual Humans Increase Willingness to Disclose. *Computers in Human Behavior* 37 (2014), 94–100.

[16] Yosi Mass, Boaz Carmeli, Haggai Roitman, and David Konopnicki. 2020. Unsupervised FAQ Retrieval with Question Generation and BERT. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. 55–60.

[17] S. Matthews and V. Yip. 2011. *Cantonese: A Comprehensive Grammar.* Routledge, New York.

[18] Clara H. McCreery, Namit Katariya, Anitha Kannan, Manish Chablani, and Xavier Amatriain. 2020. Effective Transfer Learning for Identifying Similar Questions: Matching User Questions to COVID-19 FAQs. In *Proc. KDD (Health Day Paper)*.

[19] Gustavo Penha and Claudia Hauff. 2020. What does BERT know about books, movies and music? Probing BERT for Conversational Recommendation. In *Proc. RecSys*.

[20] Dan Russell, Letitia Anne Peplau, and Mary Lund Ferguson. 1978. Developing a Measure of Loneliness. *Journal of Personality Assessment* 42, 3 (1978), 290–294.

[21] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proc. AAAI*.

[22] Oriol Vinyals and Quoc V. Le. 2015. A Neural Conversational Model. In *Proc. Deep Learning Workshop*.

[23] Taesun Whang, Dongyub Lee, Chanhee Lee, Kisu Yang, Dongsuk Oh, and Heuiseok Lim. 2020. An Effective Domain Adaptive Post-Training Method for BERT in Response Selection. In *Proc. Interspeech*.

[24] Tak-Sum Wong and John Lee. 2018. Register-sensitive Translation: A Case Study of Mandarin and Cantonese. In *Proc. Association for Machine Translation in the Americas (AMTA)*.

[25] Douglas G. Wren and Jeri Benson. 2004. Measuring Test Anxiety in Children: Scale Development and Internal Construct Validation. *Anxiety, Stress, & Coping: An international Journal* 17, 3 (2004), 227–240.

[26] Yu Wu, Wei Wu, Zhoujun Li, and Ming Zhou. 2018. Learning Matching Models with Weak Supervision for Response Selection in Retrieval-based Chatbots. In *Proc. 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

[27] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536* (2019).

[28] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *Computational Linguistics* 46, 1 (2020), 53–93.

[29] Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 1118–1127. https://doi.org/10.18653/v1/P18-1103