Restatement and Question Generation for Counsellor Chatbot

John S. Y. Lee, Baikun Liang, Haley H. M. Fong
Department of Linguistics and Translation
City University of Hong Kong
Hong Kong SAR, China
{jsylee, baikliang2, heimfong3}@cityu.edu.hk

Abstract

Amidst rising mental health needs in society, virtual agents are increasingly deployed in counselling. In order to give pertinent advice, counsellors must first gain an understanding of the issues at hand by eliciting sharing from the counsellee. It is thus important for the counsellor chatbot to encourage the user to open up and talk. One way to sustain the conversation flow is to acknowledge the counsellee's key points by restating them, or probing them further with questions. This paper applies models from two closely related NLP tasks — summarization and question generation — to restatement and question generation in the counselling context. We conducted experiments on a manually annotated dataset of Cantonese post-reply pairs on topics related to loneliness, academic anxiety and test anxiety. We obtained the best performance in both restatement and question generation by fine-tuning BertSum, a state-of-the-art summarization model, with the in-domain manual dataset augmented with a large-scale, automatically mined open-domain dataset.

1 Introduction

Advances in dialog modeling have facilitated chatbot use in many domains (Li et al., 2016; Zhou et al., 2020; Wang et al., 2020a). They are now also increasingly deployed for mental health assistance, including counselling (Fitzpatrick et al., 2017).

Dialogs in counselling share some common characteristics with those in other domains. Advice generation, for example, can be implemented with a Q&A model that retrieves counselling materials from a knowledge base (Liu et al., 2013; Huang et al., 2015). Empathetic language — words that reflect the feelings of one's interlocutors — is conducive to establishing rapport with the counsellee. Research in empathetic response generation has led to systems that can recognize the emotional state of the user, and generate responses tailored to that state (Lubis et al., 2018; Lin et al., 2019). The counsellor must also encourage the counsellee to open up and talk in order to gain an adequate understanding of the issues at hand. A common strategy to sustain the conversation flow is to use "encouragers" (Ivey and Ivey, 2003), such as back-channel phrases, restatements and questions. A good restatement acknowledges main points from the counsellee by paraphrasing or summarizing them. A helpful question elicits elaboration on a key point and invites collaborative problem solving. Table 1 shows some examples.

This paper focuses on automatic generation of restatements and questions for counselling dialogs. Specifically, it addresses two research questions:

- Text summarization and question generation are NLP tasks that are potentially relevant to the counselling domain. Can we adapt models designed for these tasks to produce high-quality restatements and questions for a counsellor chatbot?
- Dialog data for domain-specific tasks such as counselling is often limited. Can we leverage open-domain dialog data to improve restatement and question generation?

Our experiments compare a number of summarization, question generation and dialog models for the single-turn reply generation task. We obtained the strongest model by fine-tuning BertSum (Liu and Lapata, 2019), a state-of-the-art summarization model, with an in-domain, manually annotated dataset augmented with a large-scale, automatically mined open-domain dataset.

After summarizing previous work (Section 2) and presenting our dataset (Section 3), we describe our approach for restatement and question generation (Section 4). We then report experimen-

Proceedings of the 1st Workshop on NLP for Positive Impact, pages 1–7
Bangkok, Thailand (online), August 5, 2021.

Post	Restatement	Question
(a) 每逢測驗都一定會夜晚唔食飯 專心温習 同自己講 我一定唔可以輸 Before a test, I skip dinner to study and I say to myself, "I must not lose"	你一定唔可以輸 You must not lose	你同邊個比賽呀？ Who are you competing with?
(b) Professor教書教得咁廢考試又出勁難 The professor teaches poorly and gives a really hard exam	考試勁難 Exam is extremely hard	你考試係咪唔識做？ Are there questions you can't answer in the exam?
(c) 我估我到考試果陣會頭痛，我以前都試過系咁 I just knew I'll get a headache during the exam, like I did before.	你擔心考試時會頭痛 You worry you'll get a headache during the exam	你有冇試過去搵醫生睇睇呢? Have you tried to consult a doctor?
(d) 朋友真係咁易識咩...唔想要損友... Making friends is not so easy ... [I] don't want bad friends ...	你覺得唔容易識朋友 You think it's not easy to make friends	係咪覺得損友好冇益？ You think bad friends are bad for you?

Table 1: Example post-statement and post-question pairs from our manually annotated dataset (Section 3.1) addressing issued related to (a,b) academic anxiety; (c) test anxiety; and (d) loneliness

tal results (Section 5) and conclude (Section 6). Our datasets are available for download from https://github.com/CantoneseCounsellorChatbot

2 Previous work

While chatbot response generation has exploited models from machine translation (Ritter et al., 2011) and question answering (Liu et al., 2013), there has been less effort in leveraging those from other NLP tasks such as text summarization and question generation. This section reviews research in these two fields.

2.1 Text summarization

Text summarization models, which condense an input text into a shorter version, can generate short summaries or headlines (Rush et al., 2015). Pre-trained language models such as BERT (Devlin et al., 2019) have been shown to boost the quality of summarization, among many other NLP tasks. Among the best-performing models is BertSum, which uses a document-level BERT-based encoder to express the semantics of the input text document and obtain sentence representations (Liu and Lapata, 2019). Its fine-tuning schedule adopts different optimizers for the encoder and the decoder, and has been shown to improve performance by alleviating the mismatch between them.

Compared to open-domain dialogs, a human counsellor more often gives shorter replies and reflects the points made by the counsellee. Summarization models can therefore potentially be helpful in generating restatements in the counselling domain. Generic summarization models, however, likely need to be fine-tuned since restatements are not identical to summaries. In Table 1(c), for instance, the perspective changes from first person to second person ('I'll get a headache' → 'You'll get a headache'); empathetic words are also inserted to diagnose the counsellee's emotion ('You worry ...'). To our knowledge, this is the first reported evaluation on applying a summarization model to counselling dialog generation.

2.2 Question generation

A question generation model composes a question from an input text. Neural question generation algorithms have recently attained state-of-the-art performance. For example, a sequence-to-sequence model with an attention mechanism has been proposed by Du et al. (2017). Answer separation techniques have further improved question quality (Kim et al., 2019).

Question generation is slightly different in the dialog context in that the answer should generally not be found in the input text, i.e., the previous utterances, so that the question would not seem redundant. Question generation models have been deployed to engage users in a conversation (Mostafazadeh et al., 2016), but the research was focused on images. Template-based approaches, as exemplified by ELIZA (Weizenbaum, 1983), can also transform the user's statements into questions. These templates are labor-intensive to

Post-reply type	Pairs	Length	
		post	reply
Post-restatement	12,634	40.1	7.9
Post-question	9,036	36.8	11.1

Table 2: Statistics on manual dataset (average length in number of characters)

Post-reply type	Method	Pairs	Length	
			post	reply
Post-restatement	Extraction	72.6K	13.6	6.3
	Matching	36.9K	47.6	6.2
Post-question	Extraction	80.7K	12.0	6.3
	Matching	33.1K	22.8	10.9

Table 3: Statistics on automatically mined dataset (average length in number of characters)

construct, however, and may not provide sufficient coverage.

3 Data

Our data consists of *post-reply pairs*, a term that will be used henceforth to refer to both post-restatement and post-question pairs. This section describes the construction process of two datasets, which contain in-domain, manually crafted (Section 3.1) and open-domain, automatically mined (Section 3.2) post-reply pairs, respectively.

3.1 Manual dataset

We recruited 10 undergraduate students to collect Cantonese social media posts with content concerning loneliness, academic and test anxiety. For each of the 6,294 posts collected, human annotators marked a text span as their "target phrase", and composed a restatement and/or question for that phrase. As shown in Table 2, the dataset contains 12,634 post-restatement pairs and 9,036 post-question pairs. There are on average 2.2 gold restatements per post, and 1.6 gold questions per post.

3.2 Automatically mined dataset

This dataset was automatically mined from the LCCC dataset (Wang et al., 2020b), which consists of 6.8 million Mandarin dialogs; and from 89K post-reply pairs crawled from Cantonese discussion forums in Hong Kong. We used two methods to generate post-reply pairs:

Extraction. To produce post-restatement pairs, we identified the longest common string of the post and the reply in each post-reply pair in the open-domain corpora above. We extracted all pairs whose longest common string contains at least four characters, and used the repeated string in the post as the restatement. To extract post-question pairs, we identified post-reply pairs whose reply starts with a short question, defined as a question mark preceded by no more than 10 characters.

Matching. We identified all posts that contain a text span that matches a target phrase in the manual dataset (Section 3.1). We then reused the restatement and/or question for that target phrase to form a new post-restatement and/or post-question pair.

4 Approach

We first construct and evaluate models for restatement generation and for question generation separately (Section 4.1). We then combine these models to interleave restatements and questions in a counselling dialog (Section 4.2).

4.1 Restatement and Question Generation

We focus on generation-based rather than retrieval-based models, in order to tailor restatements and questions specifically to the content in the post. For each of the following approaches, we trained a restatement generation model by fine-tuning the pre-trained model with post-restatement pairs in the manual dataset (Section 3.1); we then separately trained a question generation model in a similar fashion.

DialoGPT We used *GPT2 for Chinese chitchat*[1], a dialog model that is based on DialoGPT (Zhang et al., 2020) and trained on GPT2-Chinese (Du, 2019). We fine-tuned the pre-trained model with our post-reply pairs (Section 3.1).[2]

mT5 Competitive question generation models can be built by fine-tuning the Google T5 model (Pan et al., 2021). Adopting a similar approach with mT5 (Xue et al., 2021), a multilingual variant of T5, we fine-tuned the mT5-base model with our post-reply pairs.[3]

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

[2] We used AdamW with a learning rate of 1.5e-4 and 2000 warmup steps as the optimizer. We fine-tuned the model for 50 epochs with batch size 32.

[3] We used a learning rate of 1e-4 and fine-tuned the model for 10 epochs with batch size 32, with the software provided at http://github.com/patil-suraj/question_generation

BertSum BertSum is a state-of-the-art summarization model (Liu and Lapata, 2019). We used the abstractive summarization model, which uses a standard encoder-decoder framework. The encoder is the pre-trained Bert and the decoder is a 6-layered Transformer with random initialization. We fine-tuned its pre-trained bert-base-chinese model with our post-reply pairs.[4]

Global Encoding The Global Encoding framework, which has shown competitive result in text summarization, seeks to improve the representations of the source-side information by using global information of the source context (Lin et al., 2018). Similar to above, we fine-tuned the pre-trained model with our post-reply pairs.[5]

Oracle Retrieval To gauge the maximum performance of a retrieval-based paradigm, this algorithm selects the highest-scoring reply in the training set in terms of ROUGE-L.

We further fine-tuned the DialoGPT, mT5, BertSum and Global Encoding models with the automatically mined dataset (Section 3.2). The resulting models are denoted as DialoGPT$^+$, mT5$^+$, BertSum$^+$, and Global Encoding$^+$.

4.2 Interleaving restatements and questions

A conversation becomes monotonous and even irritating if the counsellor repeatedly gives restatements or asks questions. Using DialoGPT and BertSum$^+$, the two strongest models for question generation (Table 5), we investigated the following methods to choose between a restatement candidate and question candidate as the reply.

BertSum$^+_{R+Q}$ This model is trained with the same settings as BertSum$^+$ (Section 4.1), except that it is fine-tuned with *both* post-restatement and post-question pairs.

BertSum$^+$ (threshold) This algorithm responds with a question when the BertSum$^+$ model for questions surpasses a confidence threshold; otherwise, it responds with a restatement. The tuning of the threshold will be described in Section 5.3.

BertSum$^+$ (random) This algorithm randomly chooses either the BertSum$^+$ model for restatements or the BertSum$^+$ model for questions.

BertSum$^+$ (ceiling) Designed to measure the maximum performance of BertSum$^+$, this algorithm identifies the subset of posts for which BertSum$^+$ generates the highest-scoring questions in terms of ROUGE-L. It replies to these posts with the generated questions, and to the remainder with restatements.

DialoGPT (ceiling) Same as above, the algorithm uses DialoGPT rather than BertSum$^+$.

5 Experimental results

All results are based on 5-fold cross-validation on the manual dataset (Section 3.1). Following previous research, our evaluation metrics include BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). In addition, we report results with METEOR (Banerjee and Lavie, 2005) and BertScore (Zhang et al., 2019).

5.1 Restatement generation

Table 4 shows the results for restatement generation. When fine-tuned on the manual dataset only, DialoGPT yielded a ROUGE-L score of 0.5525, outperforming Global Encoding (0.4031), mT5 (0.4960) and BertSum (0.4938).

When augmented with the automatically mined post-restatement pairs, BertSum$^+$ achieved the best ROUGE-L score (0.7142). It also outperformed other models in terms of BLEU, METEOR and BertScore. In terms of ROUGE-L, it even surpassed Oracle Retrieval (0.6932), which means that the restatements generated by the model were superior to the best available in the training set.

5.2 Question generation

Generally, automatically generated questions have lower ROUGE scores than restatements (Table 5). DialoGPT achieved only 0.4160 ROUGE, compared to 0.5525 for restatements. It outperformed both Global Encoding (0.3766) and BertSum (0.3602).

[4]We used two Adam optimizers with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for the encoder and the decoder, respectively, and learning rate $lr_E = 0.002$ and $lr_D = 0.1$. All models were trained for 200,000 steps. Model checkpoints were saved and evaluated on the validation set every 2,500 steps. We selected the best checkpoint based on their evaluation loss on the validation set.

[5]We used Adam with learning rate 0.0003 and learning rate decay parameter 0.5. We fine-tuned the model for 30 epochs with batch size 64.

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	METEOR	BertScore
DialoGPT	0.5587	0.4369	0.5525	0.5010	0.5135	0.4954
DialoGPT$^+$	0.5740	0.4656	0.5681	0.5038	0.5303	0.5127
Global Encoding	0.4114	0.2588	0.4031	0.3200	0.3347	0.3511
Global Encoding$^+$	0.6136	0.5079	0.6073	0.5449	0.5738	0.5508
mT5	0.5004	0.4133	0.4960	0.4102	0.4332	0.4276
mT5$^+$	0.5550	0.4787	0.5520	0.4751	0.5051	0.4712
BertSum	0.5013	0.3171	0.4938	0.4315	0.3986	0.3618
BertSum$^+$	**0.7184**	**0.6362**	**0.7142**	**0.6518**	**0.6881**	**0.6647**
Oracle Retrieval	0.6902	0.6011	0.6932	0.6709	0.6878	0.6604

Table 4: Model performance on restatement generation (the + superscript means the training set includes the automatically generated data)

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	METEOR	BertScore
DialoGPT	0.4252	0.2601	0.4160	0.4157	0.3605	0.4273
DialoGPT$^+$	0.3952	0.2360	0.3848	0.3803	0.3251	0.3905
Global Encoding	0.3845	0.2085	0.3766	0.3658	0.3082	0.3820
Global Encoding$^+$	0.4073	0.2516	0.3990	0.3887	0.3372	0.4004
mT5	0.3807	0.2415	0.3699	0.3669	0.3184	0.4152
mT5$^+$	0.3564	0.2293	0.3472	0.3338	0.2975	0.3932
BertSum	0.3676	0.1718	0.3602	0.3568	0.2591	0.2992
BertSum$^+$	**0.4752**	**0.3171**	**0.4665**	**0.4390**	**0.4002**	**0.4658**
Oracle Retrieval	0.6597	0.5612	0.6538	0.6401	0.6111	0.6626

Table 5: Model performance on question generation (the + superscript means the training set includes the automatically generated data)

When augmented with the automatically mined dataset, BertSum$^+$ again showed significant gains in performance. It achieved the highest ROUGE-L score (0.4665), followed by Global Encoding$^+$ (0.3990) and DialoGPT$^+$ (0.3848). Although mT5 is designed for question generation, its output scored lower than the other models in ROUGE-L, both when it is trained without (0.3699) and with the automatically mined data (0.3472).

5.3 Interleaving restatements and questions

Since it is more challenging to generate questions than restatements, a fair comparison between the algorithms requires a constant *question frequency* — i.e., the proportion of posts in the evaluation data to which the chatbot offers a question as response. The BertSum$^+_{R+Q}$ model generated questions 27.1% of the time and restatements 72.9% of the time.[6] We therefore set the confidence threshold for the BertSum$^+$ (threshold) model such that its question frequency would also be 27.1%. We likewise configured the BertSum$^+$ (random) model to randomly choose 27.1% of the posts to reply with questions.

As shown in Table 6, BertSum$^+$ (threshold) achieved the best performance at 0.7013 ROUGE-L, higher than its random counterpart (0.6730), BertSum$^+_{R+Q}$ (0.6702), as well as DialoGPT (ceiling) (0.5604). It suffered only a degradation of 0.04 in comparison to BertSum$^+$ (ceiling), which picks the optimal posts for question generation. This result suggests the effectiveness of selecting reply type with a confidence threshold.

One advantage of BertSum$^+$ (threshold) over BertSum$^+_{R+Q}$ is the ease with which question frequency can be adjusted to suit different conversation styles. Figure 1 plots its ROUGE-L score at various question frequencies. Since question generation is more difficult, the score decreases as questions are selected as the reply to a larger proportion of posts. BertSum$^+$ (threshold) outperformed both its random counterpart and DialoGPT (ceiling) at all question frequencies.

[6]The output is considered a question if it achieves a higher ROUGE-L score with the gold output in the post-question pair than the post-restatement pair (Section 3.1).

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	METEOR	BertScore
BertSum$^{+}_{R+Q}$	0.6752	0.5703	0.6702	0.6670	0.6308	0.6379
BertSum^{+} (random)	0.6793	0.5664	0.6730	0.6884	0.6376	0.6412
BertSum^{+} (threshold)	**0.7071**	**0.6061**	**0.7013**	**0.7232**	**0.6673**	**0.6621**
BertSum^{+} (ceiling)	0.7504	0.6610	0.7456	0.7548	0.7137	0.7122
DialoGPT (ceiling)	0.5679	0.4371	0.5604	0.5156	0.5111	0.5147

Table 6: Model performance on response generation of either restatement or question (the + superscript means the training set includes the automatically generated data)

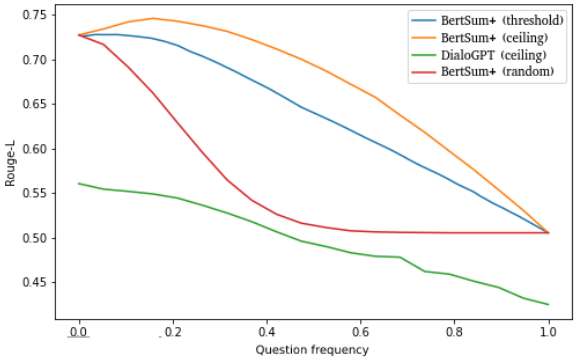

Figure 1: ROUGE-L score of BertSum^{+} (threshold), BertSum^{+} (ceiling), BertSum^{+} (random) and DialoGPT (ceiling) at various question frequencies.

6 Conclusion

Restatements and questions are common conversation strategies in counselling. This paper has investigated automatic generation of these two reply types by exploiting models of two closely related NLP tasks — summarization and question generation. We obtained the best generation performance for both reply types by fine-tuning BertSum, a state-of-the-art summarization model, with an in-domain, manually annotated dataset augmented with a large-scale, automatically mined open-domain dataset. We then showed that restatements and questions can be interleaved with a confidence score threshold.

To the best of our knowledge, this is the first reported application of summarization models on chatbot response generation in the counselling domain. It is hoped that our proposed techniques can improve the quality of a counsellor chatbot for the public. Further research is needed to take into account the progress of the counselling session when selecting a reply (Althoff et al., 2016; Zhang and Danescu-Niculescu-Mizil, 2020), and to measure correlation with counselling outcomes.

Acknowledgments

This work was 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

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

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

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to Ask: Neural Question Generation for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zeyao Du. 2019. Gpt2-chinese: Tools for training gpt2 model in chinese language. https://github.com/Morizeyao/GPT2-Chinese.

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):e19.

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*, 9085:133–145.

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

Yanghoon Kim, Hwanhee Lee, Joongbo Shin, and Kyomin Jung. 2019. Improving Neural Question Generation Using Answer Separation. In *Proc. 33rd AAAI Conference on Artificial Intelligence (AAAI-19)*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proc. ACL*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, page 74–81, Barcelona, Spain.

Junyang Lin, Xu Sun, Shuming Ma, and Qi Su. 2018. Global Encoding for Abstractive Summarization. In *Proc. ACL*.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of Empathetic Listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 121–132.

Yang Liu and Mirella Lapata. 2019. Text Summarization with Pretrained Encoders. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, page 3730–3740.

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)*, pages 67–72.

Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2018. Eliciting Positive Emotion through Affect-Sensitive Dialogue Response Generation: A Neural Network Approach. In *Proc. 32nd AAAI Conference on Artificial Intelligence (AAAI-18)*.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating Natural Questions About an Image. In *Proc. 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Unsupervised Multi-hop Question Answering by Question Generation. In *Proc. NAACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. ACL*.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-Driven Response Generation in Social Media. In *Proc. EMNLP*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proc. EMNLP*.

Xiaofeng Wang, Shuai Chen, Tao Li, Wanting Li, Yejie Zhou, Jie Zheng, Qingcai Chen, Jun Yan, and Buzhou Tang. 2020a. Depression Risk Prediction for Chinese Microblogs via Deep-Learning methods: Content Analysis. *JMIR Medical Informatics*, 8(7).

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020b. A large-scale chinese short-text conversation dataset. In *Proc. CCF International Conference on Natural Language Processing and Chinese Computing*, pages 91–103.

Joseph Weizenbaum. 1983. ELIZA–A Computer Program For the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*, 26(1):23–28.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proc. NAACL*.

Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020. Balancing Objectives in Counseling Conversations: Advancing Forwards or Looking Backwards. In *Proc. ACL*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Proc. ACL*.

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):53–93.