Empathetic Dialogue Systems

Sahand Sabour

May 2021

1 Introduction

In recent years, research on dialogue systems has gained significant attention. This is believed to be due to the high potential of these systems and the various benefits they could bring to the daily lives of their consumers once they reach an acceptable level. In general, dialogue systems are divided into task-oriented and non-task-oriented systems (also known as open-domain dialogue systems, conversational agents, and chatbots) [1].

Task-oriented systems aim to assist the user in completing a task (e.g. making a meal, buying a train ticket, booking a hotel room). Amazon's Alexa and Apple's Siri are prominent examples of such systems. On the contrary, a non-task-oriented system aims to act as an interlocutor to provide entertainment, companionship, and essentially, a natural conversation between human and machine [2]. These systems are said to be open-domain as there are no boundaries for the topics that could be discussed (i.e. the conversation is not about a single specific topic and could contain topics of all backgrounds).

The main objective of research on open-domain dialogue systems is to find a method of modeling human communicative behaviors, which enables machines to produce responses that seem both genuine and authentic. A core attribute of human conversations is the interlocutors' unique ability to express empathy towards each other. Hence, enabling dialogue systems to express empathy towards their users is necessary as they are expected to show understanding, be engaging, and provide effective companionship.

In this paper, a thorough literature review on the topic of empathetic dialogue systems including various definitions of the term *empathy*, its related concepts, and recent frameworks will be provided. Accordingly, the shortcomings in the field and propose potential solutions to these problems would be demonstrated. The detailed research plan and objectives will be provided likewise.

2 Literature Review

In this section, the proposed definitions of empathy in the psychology literature, as well as similar concepts, are introduced. In addition, recent work on empathetic dialogue systems for different tasks of classification, generation, and rewriting is reviewed. Lastly, widely-used approaches for evaluating an empathetic dialogue system are explored.

2.1 Definition of Empathy

Empathy is commonly known as a complex psychological construct, which has become the focus of much attention during the past decades [3]. Given that empathy is a fairly new term in the literature, there is no specific or widely accepted definition of empathy in the fields of social psychology and psychotherapy [4, 5, 6], which are believed to be essential fields for constructing human-like dialogue systems. Turner (2012) defined empathy as the ability that enables an individual to experience the feeling of another person [7]. Empathy has also been defined as a complex multi-dimensional construct with two broad aspects of affect and cognition [8]. In this project, we adopt the definition of empathy as the ability to perceive, understand, and react appropriately to the situation's context, the implicit emotions, and attitudes of others [9].

2.2 Related Concepts

Due to the complex nature of empathy, many closely related concepts, such as sympathy (intentionally reacting emotionally), mimpathy (mimicking another person's emotions without experiencing them), and perspective-taking (observing the situation from the perspective of another person) are commonly used in place of empathy; however, they cannot fully represent the concept of empathy and are merely parts of it [10]. Instead, empathy could be realized as a process, which could consist of any number of these related concepts [9].

In the field of Natural Language Processing (NLP), much work has used *emotion* has been widely used interchangeably with *empathy*, proposing that producing emotional responses demonstrates an empathetic agent [11, 12]. Towards this direction, Zhou et al., (2018) proposed the Emotional Chatting Machine (ECM) for generating responses based on a specific emotion [13]. Asghar et al., (2018) employed the LSTM architecture the emotional aspects to generate emotional responses [14]. Recent work also proposes utilizing word combinations [15], rationality [16], the emotional information of the user utterance [17], an external knowledge base [18], and situational context [19] to generate more appropriate responses. However, as discussed, though emotion is one of the main factors of empathy, it can not be used to fully represent empathy. Hence, generating emotional responses would not lead to an empathetic dialogue system.

2.3 Empathy in NLP

Recent work has considerably focused on implementing empathy in conversational systems and the literature demonstrates great potential for this topic. An essential part of creating an empathetic dialogue system, or any dialogue system in general, is data. In 2019, the Empathetic Dialogues dataset was released [20], which is believed to be the first conversation dataset that targeted empathy in conversations. This provided researchers with a fairly solid foundation and a benchmark for the task of empathetic response generation.

Initially, many studies focused on employing emotion as the determining factor of empathy. There were many different approaches to perform such an implementation. Rashkin et al., (2019) and Lin et al. (2020) proposed generating empathetic responses based on the detected user emotion [20, 21]. Shin et al., (2020(implemented a look-ahead approach, where the response is generated based on the predicted future state of user's emotions [22]. Lin et al. (2020) formulated this problem as the understanding of the user's emotion context [23] while Majumder et al., (2020) proposed that the generated response should mimic the user's emotion to a degree [24]. In addition, recent approaches have also considered the relative context, dialog acts, and actions of user utterances [25] as well as an external knowledge [26] to produce empathetic responses. Recently, Zheng et al., (2021) suggested that emotion may not be the only factor determining empathy and proposed utilizing dialogue acts and empathetic mechanisms as well as emotions to produce better empathetic responses [27]. Furthermore, Xie and Pu (2021) proposed training a transformer with emotion embeddings on a large-scale automatically annotated empathetic dataset [28].

However, the mentioned work does not include a target for generating empathetic responses, which makes for a rather shallow achievement: empathetic responses are generated regardless of whether empathy is needed or how much empathy is needed[29]; in addition, the conducted evaluations are not solid as existing work asks for a 1-5 score for the generated responses during the human evaluation. As mentioned, empathy is a complex concept that does not have a clear and universal definition within the fields of social psychology and psychotherapy. Since, to our knowledge, no work has been published to propose a universal definition for empathy in NLP. Hence, all the above approaches provide a definition of empathy that is most suitable for their respective objectives, which prevents comparison between different methods. Given that work on this topic is fairly new, it would be reasonable to utilize this concept to address a specific target whose goals and objectives can be observed and evaluated.

Towards utilizing empathy for achieving a specific target, Sharma et al., (2020) proposed a computational approach for measuring expressed empathy in mental health support using three empathetic communication mechanisms (i.e. exploration, interpretation, and emotional reaction) [30]. Although these mechanisms may not be sufficient to define empathy as a whole, the obtained results showed that they are a well-designed representation of empathy in mental health support. Moreover, they also proposed a new task of empathetic rewriting, which teaches mental health supporters how to be more empathetic based on the computational model of communication mechanisms [31]. Their work is admirable as it is beneficial for an individual's mental well-being, and their chosen target (empathy for mental health support) can be evaluated via observing the changes in the user's emotions upon responding, inquiring about user experience, etc.

2.4 Evaluating Empathetic Dialogue Systems

Evaluating a dialogue system is important as it enables a comparison between different methods and approaches. There are many automatic metrics for evaluating dialog systems. Originally, evaluation metrics of other generation tasks such as machine translation [32, 33] and text summarization [34] were used to evaluate dialog systems; however, these metrics are not sufficient as an input in the open-domain settings could have many possible responses.

Given that empathy is a complex term with numerous definitions, creating an automatic metric for measuring and evaluating empathy is non-trivial. Hence, human judgments are still the most viable method for evaluating dialogue systems, especially empathy. However, all of the above work on empathetic response generation employs a rather shallow human evaluation survey: they show workers a set of generated responses and ask them to assign each set an empathy score from 1-5. Since each work on this topic uses a definition of empathy that is most aligned with their proposed approach, it is evident that human evaluation results would support the effectiveness of their work if they are asked to measure empathy based on an arbitrary definition. Apart from [35], which proposed empathy as an aspect of a dialogue system's likeability (i.e. measure of how likable the generated responses are), there has been no work addressing this problem and providing a solid framework for evaluating empathy.

3 Existing Problems

There are several existing problems in the published literature of empathetic dialogue systems. Firstly, there is no large-scale high-quality dataset for empathy. Empathetic Dialogues [20] is the most commonly-used dataset for empathetic response generation; yet, conversations in this dataset seem unnatural, short, and low-quality. In addition, crowdsourcing for data collection is expensive and requires significant labor costs. Hence, the only way of obtaining a large-scale dataset is by utilizing automatic methods. However, there are no suitable annotation schemes for empathy to automatically annotate large-scale data. Though Sharma et al., (2020) provide a computational framework for expressed empathy, their proposed mechanisms are limited to mental health support (which merely covers negative empathy) and cannot be generalized.

Secondly, most of the recent work does not have a specific target for using empathy and generating empathetic responses: there is no consideration of whether empathy is needed nor how much empathy is needed when generating responses; moreover, existing work does not include any analysis on the effectiveness of the generated responses. Analyzing the effectiveness is non-trivial as there is no clear target to be achieved (i.e. it is challenging to measure success if there is not a goal to achieve nor a target to reach).

Furthermore, there are no solid evaluation methods for empathy. As mentioned, existing work conducts asks crowdsourcing workers for a score from 1-5 for empathy, which does not provide a solid judgment of the system's performance, especially for its empathetic ability. Moreover, Current automatic metrics are unable to provide a reliable measure of empathy.

4 Motivation

As mentioned, one of the main attributes of human conversations is the ability to express empathy. However, it is believed that more attention should be paid to the reason why we express empathy. As humans, we express empathy towards others to demonstrate our understanding of their situation and build trust in our relationships. We use empathy to provide support, acknowledgment, and affirmation towards others. For instance, we can use empathy to make the other person feel better when they are sad or make them appreciate sharing some good news with us when they are happy.

As mentioned by [31], there are major ethical issues and concerns when providing mental health support. In addition, it was reported that many individuals resort to online platforms to seek out support for their emotional needs, yet many individuals go unnoticed when they post on these platforms, which would likely discourage them from sharing their experiences[36]. Moreover, mental health may only cover topics that correspond to negative emotions, which does not fully represent the usage of empathy. Hence, it is believed that empathy could be utilized to provide emotional support. To our knowledge, there has been no prior work to address the problem of implementing empathy for online emotional support. Given that emotional support is a highly beneficial task and considerably less likely to include the dangers and issues concerned with mental health support, it is believed that the proposed topic is novel, beneficial, and most importantly, practical. In addition, by focusing on using empathy for providing emotional support, we have a specific target to strive towards and the effectiveness of different proposed approaches could be robustly evaluated. For instance, the user shows signs of feeling better about a topic they were sad about or they show signs of appreciation or interest in the conversation after sharing something that made them happy. Lastly, various support methods could be used to address different problems and provide a more satisfactory experience (e.g. *providing reassurance* when the user is having anxiety about an upcoming exam and *talking about one's own experience* when the user has failed a class.)

5 Importance and Impact

As mentioned, there are two types of dialogue systems: task-oriented and non-task-oriented (open domain). In task-oriented systems, the system has a clear goal; that is, the system aims to assist the user to carry out a task. This goal is well-defined and certain reward criteria could be assigned for completing each task. In this domain-specific setting, the agent could learn to as expected to obtain the highest rewards. However, in the open domain setting, each user query could have numerous possible answers; therefore, there is no way of assigning a reward to a certain answer without neglecting other possible answers. Thus, the goal of an open domain dialogue system is not to have the correct answer to each query but to achieve high user satisfaction and engagement.

When engaging with an open domain dialogue system, users may want to share their experiences, feelings, and emotions. In such a scenario, it is crucial for the system to have empathy: understand the user's situation, respond in a way that is thoughtful, genuine, engaging and most importantly without any judgment and toxicity. With the development of internet, social media, and online forums, many individuals tend to share their life experiences, current situation, and emotions on these platforms. This enables them to find people who have had similar experiences or share similar feelings, which would lead to emotional conversations that benefit their mental well-being. Hence, it becomes essential to create dialogue systems that can act as online viewers who comment on these posts and are capable of providing such emotional support (i.e. social good) [37]. This is believed to have a significant impact on the lives of individuals who are experiencing a rough going as well as people who feel lonely and need someone who they can talk with. Furthermore, implementing empathy in dialogue systems could also serve as a beneficial addition to numerous applications, including understanding customer feedback, providing more effective customer support, psychological counseling and therapy, mental health support, etc.

6 Expected Contributions

The expected contributions for this project are twofold: first, an empathetic dataset alongside a robust human evaluation for measuring empathy would be developed; second, an empathetic model for generating appropriate responses would be theoretically proposed, constructed, and thoroughly evaluated.

A reliable empathetic dataset should capture the natural dialogue between two humans. Moreover, the dialogues in this dataset should cover topics of various types of emotions (e.g. it should cover examples of individuals being happy, angry, and sad about something that happened in their lives). Lastly, it should be large-scale since this was mentioned as a major shortcoming of existing datasets.

A robust empathy evaluation scheme should assess different aspects of empathy

(such as coherence, emotional relevance of the response to the original user utterance, etc.), provide a clear definition for each of these aspects, and essentially, demonstrate a reusable and robust scale for future work on the topic of empathetic dialogue systems.
An empathetic Model should understand the user's situation based on the provided context and explore for more information if the context is not clear. Furthermore, it should identify what type and what level of empathy is needed according to the context and accordingly, generated a fitting response. The responses should seem genuine, be engaging, and avoid toxicity and judgments.

7 Research Plan

The expected plan for conducting this project is provided in the tables below:

Objective	Start Date	End Date	Status
Data collection	2021-05	2021-06	90% complete
Literature Review	2020-12	2021-07	60% complete
Human Annotation	2021-07	2021-08	Not started
Classifier for large-scale annotation	2021-07	2021-08	Not started
Experiments	2021-07	2021-09	Not started
Human Evaluation	2021-08	2021-09	Not started

Table 1: Part 1 of the Project

Objective	Start Date	End Date	Status
Literature Review	2020-12	2021-10	Not started
Theoretical Proposal	2020-09	2021-10	Not started
Experiments	2021-10	2021-12	Not started
Human Evaluation	2021-12	2022-01	Not started
Master's Thesis	2022-01	2022-04	Not started

Table 2: Part 2 of the Project

Research on this topic requires a thorough investigation of the existing literature, sufficient knowledge of the theories in the fields of psychology, psychotherapy, and human behavioral studies, critical thinking, as well as robust experimentation and evaluation. Therefore, for each of the above parts, extensive research on the existing work and relative approaches would be conducted. Accordingly, a detailed map of required objectives and their corresponding expected dates is extracted. After adopting an ideology or theory, experiments would be designed to robustly evaluated the reliability of such ideology and theory. Upon obtaining satisfactory results of a proposed theory, the experiments would be conducted on a much larger scale to create a product (e.g. a metric, dataset, or model). The experiments should be meticulously documented and the obtained results would be recorded and evaluated by human judges likewise. Lastly, a research paper for each of these parts would be written and submitted to a popular journal, depending on the time of completion. The expected schedule requires two papers as well as a master's thesis to be published during the 2021-22 academic year.

8 Tools and Implementation

Python [1] is the programming language that would be used in this project. The code for experimentation will be written using the popular deep learning framework PyTorch [2]. In addition, the famous natural language processing library Huggingface's Transformers [38] would be used to prototype, train, and test models as it allows reusing pre-trained models and convenient training, validation, and testing functions. Amazon Mechanical Turk (MTurk) [3] as well Facebook's ParlAI [39] framework would be used to create surveys and conduct the human evaluation.

9 Expected Difficulties

During the 2020-21 academic year, based on the reviewed literature, the conducted experiments, and numerous discussions with peers about the topic of empathy and empathetic dialogue systems, it was realized that tackling empathy without specifying an objective is a non-trivial task. There are many aspects related to empathy and considering all those factors when annotating the collected data or creating a model is rather complex and highly challenging. The biggest difficulty of this project is expected to be identifying key aspects that could be computationally explained and combining these aspects in a way that provides a genuine and authentic experience for the users.

[1] https://www.python.org/

[2] https://pytorch.org/

[3] https://www.mturk.com/

Apart from the theoretical and ideological issues, it might be challenging to design a clear and detailed guideline for crowdsourcing workers. Another challenge that arises in this aspect would be controlling the quality of the workers' performance and maintaining this supervision throughout the project.

10 Estimated Costs

The expected costs are provided in the table below:

#	Item	Cost (RMB)
1	PC	8,000
2	Data Annotation	100,000
3	Publication	2,000

In addition, 4-8 GPU cards would be required for running experiments and training different models for response generation. However, as GPUs are bought or rented per lab, not per person, their cost is not included in the above table.

References

[1] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35, 2017.

[2] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. Challenges in building intelligent open-domain dialog systems. *arXiv*, 1(1), 2019.

[3] Michel-Pierre Coll, Essi Viding, Markus Rütgen, Giorgia Silani, Claus Lamm, Caroline Catmur, and Geoffrey Bird. Are we really measuring empathy? Proposal for a new measurement framework. *Neuroscience Biobehavioral Reviews*, 83(October):132–139, dec 2017.

[4] David Macarov. Empathy: The charismatic chimera. *Journal of Education for Social Work*, 14(3):86–92, 1978.

[5] R Elliott, A.C Bohart, J.C Watson, and L.S Greenberg. Empathy. *Psychotherapy relationships that work (2nd ed.)*, pages 132–152, 2011.

[6] Anne M Dohrenwend. Defining empathy to better teach, measure, and understand its impact. *Academic Medicine*, 93(12):1754–1756, 2018.

[7] Stephen Turner. The Strength of Weak Empathy. *Science in Context*, 25(3):383–399, sep 2012.

[8] Mark H. Davis. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1):113–126, 1983.

[9] Sevgi Coşkun Keskin. From what isn't Empathy to Empathic Learning Process. *Procedia - Social and Behavioral Sciences*, 116:4932–4938, 2014.

[10] Benjamin MP Cuff, Sarah J Brown, Laura Taylor, and Douglas J Howat. Empathy: A review of the concept. *Emotion review*, 8(2):144–153, 2016.

[11] Xianda Zhou and William Yang Wang. MojiTalk: Generating Emotional Responses at Scale. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1:1128–1137, nov 2017.

[12] Yuzhao Mao, Qi Sun, Guang Liu, Xiaojie Wang, Weiguo Gao, Xuan Li, and Jianping Shen. Dialoguetrm: exploring the intra- And inter-modal emotional behaviors in the conversation. *arXiv*, 2020.

[13] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. *arXiv*, pages 730–738, apr 2017.

[14] Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. Affective Neural Response Generation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10772 LNCS, pages 154–166. sep 2018.

[15] Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. Generating Responses with a Specific Emotion in Dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3695, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics.

[16] Peixiang Zhong, Di Wang, Pengfei Li, Chen Zhang, Hao Wang, and Chunyan Miao. CARE: Commonsense-Aware Emotional Response Generation with Latent Concepts. *arXiv*, dec 2020.

[17] W. E.I. Wei, L. I.U. Jiayi, M. A.O. Xianling, G. U.O. Guibing, Z. H.U. Feida, Pan Zhou, H. U. Yuchong, and Shanshan Feng. Target guided emotion aware chat machine. *arXiv*, x(x), 2020.

[18] Peixiang Zhong, Di Wang, and Chunyan Miao. Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. 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)*, number 2017, pages 165–176, Stroudsburg, PA, USA, sep 2019. Association for Computational Linguistics.

[19] Sashank Santhanam and Samira Shaikh. Emotional Neural Language Generation Grounded in Situational Contexts. *arXiv*, nov 2019.

[20] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y. Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 5370–5381, 2019.

[21] Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. Caire: An end-to-end empathetic chatbot. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13622–13623, 2020.

[22] Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. Generating Empathetic Responses by Looking Ahead the User's Sentiment. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2020-May:7989–7993, 2020.

[23] Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. MOEL: Mixture of empathetic listeners. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 121–132, 2020.

[24] Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. MIME: MIMicking Emotions for Empathetic Response Generation. 2020.

[25] Rohola Zandie and Mohammad H. Mahoor. EmpTransfo: A Multi-head Transformer Architecture for Creating Empathetic Dialog Systems. 2020.

[26] Qintong Li, Hongshen Chen, Zhaochun Ren, Zhumin Chen, Zhaopeng Tu, and Jun Ma. EmpGAN: Multi-resolution Interactive Empathetic Dialogue Generation. (2018), 2019.

[27] Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. CoMAE: A Multi-factor Hierarchical Framework for Empathetic Response Generation. (Section 3), 2021.

[28] Yubo Xie and Pearl Pu. Generating Empathetic Responses with a Large Scale Dialog Dataset. 2021.

[29] Zixiu Wu, Rim Helaoui, Vivek Kumar, Diego Reforgiato Recupero, and Daniele Riboni. Towards detecting need for empathetic response in motivational interviewing. *Companion Publication of the 2020 International Conference on Multimodal Interaction*, 2020.

[30] Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*, 2020.

[31] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. *arXiv preprint arXiv:2101.07714*, 2021.

[32] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, volume 371, page 311, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[33] Alon Lavie and A Agarwal. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. *Proceedings of*

the Second Workshop on Statistical Machine Translation, 0(June):228–231, 2007.

[34] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. *Proceedings of ACL workshop on Text Summarization Branches Out*, page 10, 2004.

[35] Vitou Phy, Yang Zhao, and Akiko Aizawa. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. *Proceedings of the 28th International Conference on Computational Linguistics*, 2020.

[36] Zijian Wang and David Jurgens. It's going to be okay: Measuring Access to Support in Online Communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 33–45, Stroudsburg, PA, USA, 2018. Association for Computational Linguistics.

[37] Peng Qi, Jing Huang, Youzheng Wu, Xiaodong He, and Bowen Zhou. Conversational AI Systems for Social Good: Opportunities and Challenges. 2021.

[38] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[39] A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*, 2017.