

# An Approach for Internet Meme Incorporated Open-domain Dialog

Haotian Wang, Zhenzhe Ying, Changhua Meng, Xiaofeng Wu, Jinzhen Lin, Lanqing Xue

Ant Group

{wht209287, zhenzhe.yzz, changhua.mch, congyu.wxf, jinzhen.ljz, lanqing.xlq}@antgroup.com

## Abstract

Internet memes are popular among people and play an important role in Internet conversations. Therefore, incorporating memes into open-domain dialog is essential. However, due to the huge number of memes and different styles, directly extracting memes features from images and passing them to a model requires huge training data to ensure the generalization of the model. At the same time, it is difficult to integrate the dialogue text with the memes information well. In this paper, we propose a new method to overcome the problems. First, we use the title of the memes instead of the image as the input of the model, and fusion the information of dialogue text and memes title through pre-training. Then reproduce historical information to avoid model forgetting and improve robustness through the model ensemble. We test our method in the Tenth Dialogue System Technology Challenge (DSTC-10) - Track1 MOD: Internet Meme Incorporated Open-domain Dialog. Our model ranks the top 1 in the hard version test dataset of task 3 and ranks the top 2 in tasks 1 and 2.

## 1 Introduction

Internet memes are widely used among people in online chat and social media. Memes vividly express our emotions, convey our attitudes, and create a lively atmosphere during conversations. However, existing chatbots only chat with pure texts in open-domain dialogues, which easily leads to a serious conversation. Therefore, properly incorporating memes into dialogues is worth exploring.

In the tenth edition of the Dialog System Technology Challenges 10 (DSTC-10), the track of “MOD: Internet Meme Incorporated Open-domain Dialog” proposes such a challenge, where the formation of utterances can be text-only, meme-only, or mixed information. Compared with traditional pure text chat, chat combined with Internet memes is a more general paradigm, and it is becoming more and more popular. Figure 1 shows two chat conversations using Internet memes, where the red part represents the emotions of the current user. It can be seen that after combining the memes, it is more convenient and straightforward to convey the current user’s emotions. Compared with the traditional text-only dialogue, the chat with the Internet memes has a higher challenge, because it needs to understand the content



Figure 1: Two illustrations of Internet meme incorporated dialogue. Dialogue system generates a response in meme-only, text-only, or a combination of them, automatically. Different emotion is annotated for each Internet meme usage in red.

and the emotion behind the memes and conduct follow-up chats based on this information.

This track is divided into three subtasks: 1) text response modeling, which evaluates the quality of text-only response, 2) meme retrieval, which discriminates the suitability of meme usage, and 3) emotion classification, which predicts the current emotion type for speakers. For these tasks, we chose PLATO-2 as our backbone model. Due to the small number of memes provided in the training dataset, we use the title of the memes instead of the image as the input of the model, and then fusion the information of dialogue text and memes title through pre-training. In the downstream task, for task1, nuclear sampling is used to generate diverse responses. For task2 and task3, we reproduce historical information to avoid model forgetting and improve robustness through the model ensemble.

## 2 Related Work

**Multi-turn dialogue model** Since the emergence of transformer-based models (Vaswani et al. 2017; Devlin et al. 2019; Radford et al. 2019), pre-training using large-scale

corpus and fine-tuning downstream tasks based on task datasets have achieved great success in natural language processing, especially in open-domain dialogue generation. DialoGPT (Zhang et al. 2020) utilizes Reddit comments to train the model to generate responses. Meena (Adiwardana et al. 2020) employs public domain social media conversations in the training process to make the chatbot more human-like. To mitigate undesirable toxic or biased traits of large corpora, Blender (Roller et al. 2021) utilize Reddit, ConvAI2, Wizard of Wikipedia, Blended Skill Talk datasets to emphasize desirable conversational skills of engagingness, knowledge, empathy, and personality. Furthermore, Plato (Bao et al. 2021) introduces discrete latent variables to tackle the inherent one-to-many mapping problem to improve response quality and explore effective training via curriculum learning (Bengio et al. 2009).

**Multimodal language model** Currently, there are many tasks of text combined with images, such as Visual Question Answering (VQA) (Antol et al. 2015; Gao et al. 2019; Li et al. 2019) and Visual dialog (Das et al. 2017; Jain, Lazebnik, and Schwing 2018; Lu et al. 2017). These tasks are usually based on one picture, extracting the information from the picture to answer or dialogue. In this challenge, there are multiple Internet memes in dialogue, and what we need to pay attention to is the emotions conveyed in the Internet meme. In addition, there are some studies on emoji prediction. For example, (Barbieri et al. 2018) use a multimodal approach to recommend emojis based on the text and images in an Instagram post. However, the number of emojis is usually relatively small, while the number of Internet memes is huge and there are various types. The most similar work to this challenge is (Laddha et al. 2020) (Gao et al. 2020), (Laddha et al. 2020) first predict the text of the next sentence based on the dialogue, and then match a similar meme based on the text. (Gao et al. 2020) employ a deep interaction network to conduct matching between candidate meme and each utterance in dialog context, then use fusion network to capture the short and long dependency of the interaction results of each utterance simultaneously. These methods mainly aim at selecting a meme rather than incorporating the meme into the dialogue input.

### 3 Dataset and Task

#### 3.1 Dataset Description

The Internet memes incorporated open-domain dialogues in the datasets are collected from the WeChat meme database. There are a total of 307 memes, including 123 GIFs. There are 52 kinds of emotions, including happiness, doubts, helplessness, distress and so on. The data statistics of the training dataset are provided in Table 1. The number distribution of memes and emotions in the training dataset is displayed in Figure 2.

The test dataset includes an easy version and a hard version which the hard version will show some memes that have not appeared in the training dataset to verify the robustness of the model to unseen memes.

Dataset Statistics	size
dialogues	66,219
utterances	927,331
tokens	7,392
Avg. utterances in a dialogue	14.0
Avg. Internet meme in a dialogue	3.76
Avg. tokens in an utterance	17.8

Table 1: Training dataset statistics.

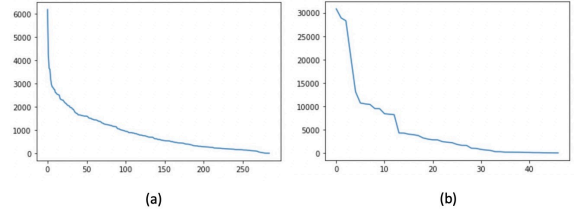


Figure 2: The number distribution of Internet memes and emotions in the training dataset after descending sorting, (a) represents Internet memes and (b) represents emotions

#### 3.2 Task Description

Overall, the Internet meme incorporated open-domain dialogue task can be formulated as: build a dialogue system through historical chat records with Internet meme to realize functions like text response modeling, meme retrieval and emotion classification.

**Text Response Modeling** The goal of this task is to generate a coherent and natural text response given the multi-modal history context, the input is multi-modal dialogue history  $(u_1, u_2, \dots, u_{t-1})$ , where  $u_i = (S_i, m_i)$ , and  $S_i$  represent text-only response and  $m_i$  represents suitable meme id.

**Meme Retrieval** The goal of this task is to select a suitable Internet meme from candidates given the multi-modal history context and generated text response, the input is multi-modal dialogue history  $(u_1, u_2, \dots, u_{t-1})$  and generated text response  $S_t$ , where  $u_i = (S_i, m_i)$ , and  $S_i$  represent text-only response and  $m_i$  represents suitable meme id.

**Emotion Classification** The goal of this task is to predict the emotion type when respond with an Internet meme, the input is multi-modal dialogue history  $(u_1, u_2, \dots, u_t)$ , where  $u_i = (S_i, m_i)$ , and  $S_i$  represent text-only response and  $m_i$  represents suitable meme id.

### 4 Approach

The main goal of this challenge is how to integrate Internet memes into the multi-turn dialogue while at the same time adapting the model to unseen memes. We first try to input the image information of the meme into the model directly. The benefits of this method are obvious, there will be no loss of information, and theoretically, it can be adapted to unseen memes. However, in the process of trying, we found that this method has two shortcomings: 1. Among the 307 Internet

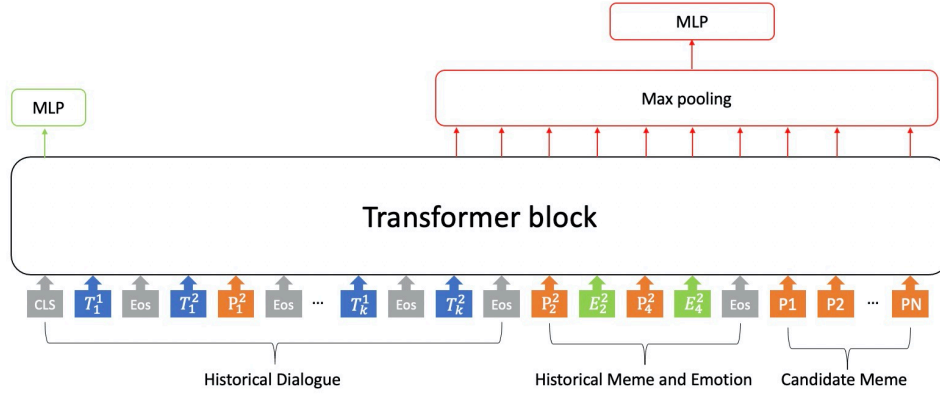


Figure 3: Meme retrieval model structure

memes, 123 are GIFs, and many GIFs cannot understand the information expressed in a single frame; 2. There are many series of Internet memes. The meme of the same series is similar in global view but different in local details, while different series vary greatly, as shown in Figure 4. However, since there are only 307 memes in the training dataset, the model cannot extract the real vital parts from the memes of different series. Therefore, the model will over-fit seriously. We tried to train the model with a part of the memes, remaining memes that have not been seen perform poorly. When analyzing the data, we found that all Internet memes have a title with the same semantics as the image of the meme. Therefore, we ended up inputting the title of the meme instead of the image into the model. For memes with the same title, we add numbers after the title to distinguish. Because the title of the meme belongs to the text, even if it is a meme that has not appeared before, it can be inferred based on the semantic information of the meme title.



Figure 4: Two different series of Internet memes.

In this challenge, we chose PLATO-2 as our backbone model. The model is based on the Unified-Transformer structure (Dong et al. 2019). It contains 1.6 billion parameters and uses 1.2 billion Chinese open-domain multi-turn dialogue datasets for training. It also introduces discrete latent variables to tackle the inherent one-to-many mapping problem. In this way, it improves response quality. Furthermore, it explores effective training via curriculum learning. Since the model did not use Internet meme information during pre-training, we continue pre-training on the data of this challenge to instruct the model to learn this information.

#### 4.1 Pre-training

Different from pre-training with BERT, we add an extra loss related to memes. The input is processed in two ways: 1). For text, we employ a masked language model (MLM) to maintain the capacity of distributed representation. 2). For Internet memes, if there is no meme in a certain round, a meme is randomly inserted; if the meme exists, there is a 15% probability that it will be replaced with a random meme. The input format and model structure is provided in Figure 5, where  $T_j^i$  refers to the text replied by the  $i$ -th user in the  $j$ -th round,  $P$  refers to the original meme,  $\hat{P}$  refers to the inserted meme, and  $\tilde{P}$  refers to replaced meme.

The MLM loss and MEME loss is defined as:

$$\mathcal{L}_{MLM} = -\mathbb{E} \sum_{m \in M} \log p(t_m | c) \quad (1)$$

$$\mathcal{L}_{MEME} = -\mathbb{E} \sum_{p \in P} \log p(l_p | c) \quad (2)$$

where  $c$  refers the all input,  $\{t_m\}_{m \in M}$  stands for masked text tokens and  $\{l_p\}_{p \in P}$  stands for Internet memes label. When the meme is the original meme,  $l_p$  is 1, otherwise 0.

To sum up, the objective of the pre-training is to minimize the following integrated loss:

$$\mathcal{L} = \mathcal{L}_{MLM} + \mathcal{L}_{MEME} \quad (3)$$

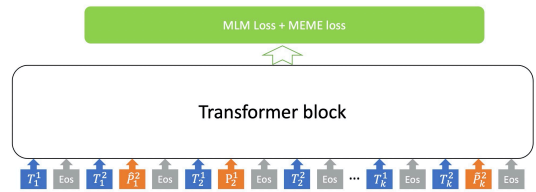


Figure 5: Pre-training structure

	task1 (easy)		task1 (hard)		task2 (easy)		task2 (hard)		task3 (easy)	task3 (hard)
	BLEU-2	DIST-1	BLEU-2	DIST-1	$R_{10}@1$	MAP	$R_{10}@1$	MAP	Acc@1	Acc@1
team 1	<b>5.08</b>	1.9	<b>5.04</b>	1.10	34.2	52.3	26.8	42.3	Nan	Nan
team 2	3.57	1.93	3.65	1.17	<b>56.8</b>	<b>72.0</b>	<b>42.0</b>	<b>58.8</b>	<b>62.3</b>	27.3
team 3	3.70	2.0	3.52	1.00	33.5	50.3	27.5	50.3	54.5	26.5
team 4	3.54	1.85	3.30	1.23	34.0	51.0	25.7	40.9	57.3	27.0
<b>team 5 (ours)</b>	3.78	<b>2.2</b>	4.03	<b>1.36</b>	34.4	52.3	27.9	45.1	58.3	<b>29.7</b>

Table 2: Results of all task

## 4.2 Task Fine-tuning

**Text Response Modeling** During training, split the entire conversation into multiple turns to form multiple samples. After the model training is completed, the predicted token of each position needs to be sampled in turn to form a complete response. Since beam search tends to produce shorter output, we use nucleus sampling (Holtzman et al. 2020) to ensure the diversity and information richness of the response.

**Meme Retrieval** In this task, the goal is to select the most suitable meme from 11 candidate Internet memes. To prevent the model from forgetting the historical information, besides the memes that appeared in the historical dialogue, we also spliced all the historical memes and emotions of the current user behind the dialogue. Furthermore, to reduce the learning difficulty, we randomly constructed 11 candidate meme sets during the training process and then spliced them to the end of the input to guide the model. In order to improve the robustness of the model, we trained models of two structures for fusion. As shown in Figure 3, the green and red frames respectively represent the upper structure of the two models and use the mean of the predicted results of the two models as the final result.

**Emotion Classification** Except that there is no candidate set, the overall model framework of the emotion classification task is the same as the meme retrieval model.

## 5 Experiments

### 5.1 Experiment Settings

The parameters of the three tasks are the same. the maximum length of input is 512 and the optimizer is Adam (Kingma and Ba 2015), with learning rate= $1e-5$ ,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ . Employ linear scheduler, which warm-up steps are 4000. There is no fixed batch size. First, randomly select 65536 samples as a pool, sort the samples by their length. Then take out B samples in turn under the premise of  $B \times L < 8192$ , where L is the maximum length of the B samples.

### 5.2 Metrics

For task 1, there are two metrics: automated evaluation and human evaluation. Automatic evaluation indicators are BLEU-2, BLEU-4, Dist-1, Dist-2. BLEU2-4 is the English word level scoring, and dist1-2 is the diversity of dialogue content at the English word level. The final ranking will be based on human evaluation results only for selected systems according to automated evaluation scores. It will address the

following aspects: grammatical/semantical correctness, naturalness, appropriateness, informativeness and relevance to given multimodal history.

For task 2, The metrics are Recall\_n@k, MAP. Recall\_n@K measures if the positive meme is ranked in the top k positions of n candidates. Mean average precision (MAP) considers the rank order. The final ranking will be based on the average score of Recall\_10@1 and MAP.

For task 3, The metric is Accuracy@K. indicates that if the correct emotion type is the highest k-class score emotion type. The final ranking will be based on the score of Accuracy@1.

## 5.3 Experiment Results

The final results of the easy version test dataset and the hard version test dataset of the three tasks are shown in Table 2. Our model ranked 1<sup>st</sup> on the hard version test dataset of task3 and ranked 2<sup>nd</sup> in the remaining tasks. Through ablation experiments on the validation dataset, we can verify the rationality of each module.

	Acc@1	Acc@3	Acc@5
model w/o pre-training	0.588	0.727	0.776
model w/o history	0.583	0.724	0.770
model w/o ensemble	0.586	0.726	0.772
model	<b>0.591</b>	<b>0.731</b>	<b>0.781</b>

Table 3: Ablation experiments.

Figure 6 and Figure 7 are two illustrations of incorrect predictions in task2 and task3 respectively. In Figure 6, mis-predicted Internet memes are also consistent with the current context. In Figure 7, for the left illustration, the emotion should be envy according to the context of the text, but the emotion of Internet meme is more like shy, which leads to the model prediction error, indicating that the model cannot integrate the information well when the meaning of the text and the meme is quite different. For the right example, the model’s prediction of surprise is also in line with the current context.

## 6 Conclusion

With the increasing popularity of Internet memes in online chat, it can be expected that Internet memes incorporated open-domain dialogue will be an important research direction in the future. In this paper, we have presented our work



<ul style="list-style-type: none"> <li>• User1: 你不要耍无赖好吧群众的眼晴都是雪亮滴！见不得沙子 (Don't be a rogue, the eyes of the masses are discerning! Can't tolerate sand)</li> <li>• User2: 这里有群众吗。。 (Are there any masses here..)</li> <li>• User1: 我不就是的吗？你不是吗？ (Am I not the masses? Aren't you the masses?)</li> </ul>	<ul style="list-style-type: none"> <li>• User1: 我电脑还好，手机放着在充电。等下用手机来 (My computer is okay, the phone is charging, and I will use it later)</li> <li>• User2: 我也在充电！我来重启电脑 (I am also charging! I'll restart the computer)</li> <li>• User1: 噢，我现在出门不带手机的 (Well, I don't bring my mobile phone when I go out now)</li> <li>• User2: 我现在发现不喜欢接电话啊最喜欢窝在家里我电脑重启好慢啊 (I now find that I don't like answering the phone. I like staying at home. My computer restarts so slowly.)</li> </ul>
 <p>Label Predict</p>	 <p>label predict</p>

Figure 6: Two illustrations of incorrect predictions in task 2.

<ul style="list-style-type: none"> <li>• User1: 昨天七夕，晚上下雨了～～ (Yesterday, Tanabata, it rained at night～)</li> <li>• User2: 我们这没下，热死了都 (It's not raining here, it's so hot)</li> <li>• User1: 我们也有热的时候。。。 (We also have hot moments..)</li> <li>• User2: 我还是觉得，你们那好 (I still think your place is better)</li> </ul>	<ul style="list-style-type: none"> <li>• User1: 朋友叫自驾走沪沽湖。现在我都还没决定去不去。你们去要多久？ (My friend asked to drive to Lug Lake by car. Now I haven't decided whether to go or not. How long are you going to play)</li> <li>• User2: 也就十天八天的。没假了 (It's only about ten days. Holidays are running out)</li> <li>• User1: 还少了呀？ (Is this short?)</li> </ul>
 <p>Label: Envy Predict: Shy</p>	 <p>Label: Doubt Predict: Surprise</p>

Figure 7: Two illustrations of incorrect predictions in task 3.

for the “MOD: Internet Meme Incorporated Open-domain Dialog” track at the Tenth Dialogue System Technology Challenge (DSTC-10). There are many series of Internet memes. The meme of the same series is similar in global view but different in local details, while different series vary greatly. Because the number of memes in the current dataset is too small, the model is unable to extract the real key parts from the memes of different series, which leads to poor performance on unseen memes. Therefore, we finally input the title of the meme instead of the image into the model and guide model learning the information through pre-training. In downstream tasks, for task1, nucleus sampling is used to generate diverse responses. For task2 and task3, we spliced all the historical memes and emotions of the current user behind the dialogue and improve robustness through the model ensemble. Especially, for task2, we additionally construct a candidate set of memes during the training process to reduce the learning difficulty of the model. Finally, We won first place in the hard version test dataset of task3 and ranked 2<sup>nd</sup> in the remaining tasks.

## References

- Adiwardana, D.; Luong, M.; So, D. R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; and Le, Q. V. 2020. Towards a Human-like Open-Domain Chatbot. *CoRR*, abs/2001.09977.
- Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.
- Bao, S.; He, H.; Wang, F.; Wu, H.; Wang, H.; Wu, W.; Guo, Z.; Liu, Z.; and Xu, X. 2021. PLATO-2: Towards Building an Open-Domain Chatbot via Curriculum Learning. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, 2513–2525. Association for Computational Linguistics.
- Barbieri, F.; Ballesteros, M.; Ronzano, F.; and Saggion, H. 2018. Multimodal Emoji Prediction. In Walker, M. A.; Ji, H.; and Stent, A., eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, 679–686. Association for Computational Linguistics.
- Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.
- Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M.; Parikh, D.; and Batra, D. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 326–335.
- Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.
- Dong, L.; Yang, N.; Wang, W.; Wei, F.; Liu, X.; Wang, Y.; Gao, J.; Zhou, M.; and Hon, H. 2019. Unified Language Model Pre-training for Natural Language Understanding and Generation. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d’Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 13042–13054.
- Gao, P.; You, H.; Zhang, Z.; Wang, X.; and Li, H. 2019. Multi-Modality Latent Interaction Network for Visual Question Answering. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 5824–5834. IEEE.
- Gao, S.; Chen, X.; Liu, C.; Liu, L.; Zhao, D.; and Yan, R. 2020. Learning to Respond with Stickers: A Framework of



- Unifying Multi-Modality in Multi-Turn Dialog. In Huang, Y.; King, I.; Liu, T.; and van Steen, M., eds., *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, 1138–1148. ACM / IW3C2.
- Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2020. The Curious Case of Neural Text Degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Jain, U.; Lazebnik, S.; and Schwing, A. G. 2018. Two Can Play This Game: Visual Dialog With Discriminative Question Generation and Answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 5754–5763. Computer Vision Foundation / IEEE Computer Society.
- Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Laddha, A.; Hanoosh, M.; Mukherjee, D.; Patwa, P.; and Narang, A. 2020. Understanding Chat Messages for Sticker Recommendation in Messaging Apps. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 13156–13163. AAAI Press.
- Li, X.; Song, J.; Gao, L.; Liu, X.; Huang, W.; He, X.; and Gan, C. 2019. Beyond RNNs: Positional Self-Attention with Co-Attention for Video Question Answering. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, 8658–8665. AAAI Press.
- Lu, J.; Kannan, A.; Yang, J.; Parikh, D.; and Batra, D. 2017. Best of Both Worlds: Transferring Knowledge from Discriminative Learning to a Generative Visual Dialog Model. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 314–324.
- Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.
- Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Smith, E. M.; Boureau, Y.; and Weston, J. 2021. Recipes for Building an Open-Domain Chatbot. In Merlo, P.; Tiedemann, J.; and Tsarfaty, R., eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, 300–325. Association for Computational Linguistics.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008.
- Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In Celikyilmaz, A.; and Wen, T., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, 270–278. Association for Computational Linguistics.