

Summary for DSTC10-MOD Track: Internet Meme Incorporated Open-domain Dialogue

<https://openai.weixin.qq.com/dstc>

Zhengcong Fei^{1,2}, Zekang Li^{1,2}, Jinchao Zhang^{3*}, Yang Feng^{1,2}, Jie Zhou³

¹ Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)

² University of Chinese Academy of Sciences

³ Pattern Recognition Center, WeChat AI, Tencent Inc, China

{lizekang19g, feizhengcong, fengyang}@ict.ac.cn

{dayerzhang, withtomzhou}@tencent.com

Track Overview

Internet memes have become one of the most important approaches for expression and emotions in social media and messaging communication (Wang et al. 2019; Beskow, Kumar, and Carley 2020; Chen 2020). Meme, which is a type of content that features a visual format of images, GIF, or short videos, can inject humor into conversations and create an emotional context (Posey et al. 2010). Compared to emojis which is limited in variety and size, memes are more expressive and engaging. Although there is an increasing interest for chatbots that can converse using multiple modalities with humans (Das et al. 2017; AlAmri et al. 2019), incorporating contextualized internet memes into multi-turn open dialogues under different situations is still under explored. This challenge aims to deal with a new task – Meme incorporated Open Dialogue (MOD), where models are required to generate a vivid response in text-only, meme-only, or mixed information, provided with a multimodal dialogue context, as shown in Figure 1. Both the form of utterances in historical context and response could be either a text only or an internet meme only or containing both text and meme, which can be considered as a more general paradigm compared with conventional text-only conversions. The MOD task is much more challenging since it requires the model to understand the multi-modal elements as well as the emotions behind them. There are three main tasks as introduced in (Fei et al. 2021): text response modeling, meme retrieval, and meme emotion classification, as listed in Table 1. The data and baseline system are publicly available¹.

Task and Data

Meme incorporated Open-domain Dialogue task

Participants are expected to build multi-modal dialogue systems based on the MOD dataset. Provided with the dia-

*Jinchao Zhang is the corresponding author.
Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹<https://github.com/lizekang/DSTC10-MOD>





<p>User1: 给我找个工作吧? (Help me find a job?) anticipate</p>  <p>User2: 什么工作? (What jobs?)</p> <p>User1: 失业了 (I'm out of work) sad</p>  <p>User2: 你准备找什么工作? (What kind of jobs are you looking for?)</p> <p>User1: 不知道, 能干成就行 (I do not know, anything is ok)</p>	<p>User1: 明天要考试好紧张, 怎么办? (I'm so nervous about the exam tomorrow. What should I do?) nervous</p>  <p>User2: 逢考必过 (Pass every exam)</p> <p>User1: 万一要是考不过呢? (What if I fail?)</p> <p>User2: 怎么可能呢? (How is that possible?)</p> <p>User1: 没有啊, 这是可能的 (No, it's possible)</p> <p>User2: 相信自己 (Believe in yourself) encourage</p> 
--	--

Figure 1: Illustrations of meme incorporated open-domain dialogues. Both history and response can be in the form of text-only, meme-only, or a combination of both. Corresponding emotion is annotated for each used Internet meme in red.

logue history consisting of utterances filled with Internet memes, the dialogue system aims to build an interesting response in the form of text-only, meme-only, or a mixed category of both. Formally, we use $U_N = [u_1, \dots, u_N]$ to denote a N turns of Internet meme incorporated dialogue, where utterance $u_i = \langle S_i, m_i \rangle$ includes the text context S_i and Internet meme m_i pair. Note that $m_i = \text{None}$ denotes there is no Internet meme incorporated in i -th utterance and $S_i = \{[\text{bos}], [\text{eos}]\}$ denotes that no text is generated as the response. Therefore, the goal of MOD is to predict the target

Task #1	Text Response Modeling
Goal	To generate a coherent and natural text response given the multi-modal history context
Input	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
Output	Natural, fluent, and informative machine text response S_t in line with dialogue history
Task #2	Meme Retrieval
Goal	To select a suitable internet meme from candidates given the multi-modal history context and generated text response
Input	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
Output	Suitable and vivid internet meme m_t in line with dialogue history
Task #3	Meme Emotion Classification
Goal	To predict the emotion type when respond with an internet meme
Input	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
Output	Emotion type c_t for current meme usage

Table 1: Summary of DSTC10 Track 1 tasks.

response \hat{u}_i for the given dialogue history \mathbf{U}_{i-1} as:

$$\hat{u}_i = \arg \max_{\langle S_i, m_i \rangle} p(\langle S_i, m_i \rangle | \mathbf{U}_{i-1}) \quad (1)$$

We further split the current scope of MOD into the following three tasks as shown in Table 1: (1) **Text Response Modeling**: given the multimodal history context \mathbf{U}_{i-1} , the task aims to generate a coherent and natural text response S_i . (2) **Meme Retrieval**: given a multimodal historical context \mathbf{U}_{i-1} and generated text response S_i , the goal here is to select a suitable meme m_i as feedback. (3) **Meme Emotion Classification**: given the multimodal 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, the goal is to predict the emotion type when respond with an internet meme.

Data Collection

Step 1: Pre-processing. For Internet meme sets, the meme candidates are firstly collected from the Internet and then chosen carefully by annotators to maintain good quality. In addition, if textual information appears in the selected Internet meme content, we will also annotate it manually. To avoid the model only utilizing the textual information and ignoring visual features, we control the proportion of memes without appeared texts in the final set to 40%. Meanwhile, to avoid multiple appropriate memes being selected under one dialogue condition, we filter out the memes with highly similar or duplicate semantic content. Finally, we obtain a total of 307 Internet memes for the subsequent data

Split	# dialogs	# sentences	# memes
Train	66,219	927,331	274
Valid	1,000	13,666	274
Easy test	1,000	10,398	274
Hard test	2,183	29,046	307

Table 2: Statistics of the Track 1 data sets.

annotating process. To facilitate the arrangement and annotating process, the Internet meme set is further split into four groups: *atmosphere adjustment*, *basic expression*, *basic emotion*, and *common semantics*, respectively.

For the initial conversation set, considering that the open-domain Internet meme is too scarce in scale, it is costly and time-consuming to collect multi-turn conversations from scratch. Thus, our annotation is based on an existing large-scale Chinese dialogue dataset with its large version (Wang et al. 2020). To make each chatting session contain rich information, we remove the dialogues which have less than 10 utterances.

Step 2: Internet meme incorporated response construction. The annotators, who are well-educated and familiar with dialogue research, are tasked to take two operations using the prepared Internet meme candidates: use one most suitable Internet meme to replace part of the text conversation or insert an Internet meme into the utterance to enhance the emotion of the current dialogues. In particular, we also ask annotators to label the emotional states when utilizing the current Internet memes. The annotators are specially instructed based on the following criteria: (i) behave naturally, and the meme usage is in line with real daily chats, (ii) the number of different Internet memes in the dataset is kept balanced to avoid meaningless gatherings and biased data. Those dialogues without any labeled Internet memes will be abandoned in the later data processing stage. Note that different from previous works, our annotation procedure is conducted posteriorly so that it will not interfere with human conversations, *e.g.*, prompting them to an overused Internet meme.

Step 3: Quality control. Before formal annotation, annotators are asked to annotate training samples until their results pass our examination. During the annotation, to eliminate the subjective inconsistency and make the annotation reliable, several specialized workers consistently monitor the collected dialogue data and perform a periodic quality check on samples. After the checking, we sample 10% data and manually check the samples ourselves. If errors are found in an annotation batch, we ask corresponding annotators to self-check and re-annotate the whole batch. In light of the above, the annotation results are closed to real-world natural conversations.

Dataset Statistics. The total detailed statistics of the MOD dataset are summarized in Table 2. The split was based on dialogues, not based on source-target pairs. MOD dataset has an average of 13.93 turns, and each turn contains

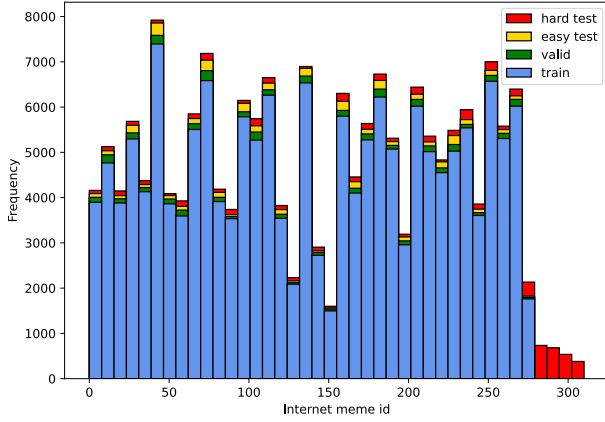


Figure 2: Internet meme frequency in the dataset. The meme usage balances without significant bias. Meme ids greater than 274 only occur in hard test set.

11.6 tokens. The text is tokenized by Chinese BERT tokenizer (Wang et al. 2020) and the vocabulary size is 13,086. We also plot the usage frequency of Internet memes and corresponding emotion in Figures 2 and 3, respectively. Each conversation contains about 4.1 Internet memes on average and each utterance equipped with an Internet meme is annotated with the corresponding emotion. Although the dialogue system is evaluated under MOD, participants can leverage any public datasets and pre-trained models to build models. In the evaluation phase, we released the test set that is divided into easy test version for all internet meme seen in the training set and hard test version for some unseen internet memes. The motivation to build a hard version is to evaluate whether a MOD model is able to transfer to exploit new internet memes. In particular, to increase the impact of MOD challenge, we also provide English version of MOD dataset under strict human-annotated quality control. We first use the machine translation engine of Google, Baidu and Wechat, to get the preliminary translation results, and then require a professional annotation team to score and calibrate the results. Finally, the selected the sentences with the highest score are served as the corresponding translation.

Evaluation Criteria

Each participating team submitted up to five system outputs each of which contains the results for all three tasks on the two unlabeled test sets. We first evaluated each submission using the automatic task-specific objective metrics as show in Table 4 by comparing to the ground-truth labels and responses. In detail, BLEU2-4 is the word level scoring, and dist1-2 is the automatic indicator of the diversity of dialogue content at the char level in Chinese and the word level in English. $\text{Recall}_n@k$ measures if the positive meme is ranked in the top k positions of n candidates. Mean average precision (MAP) consider the rank order. The top k accuracy, referred to as $\text{Accuracy}@k$, indicates that if the correct emotion type in the highest k -class score emotion type, the score of metric is 1.

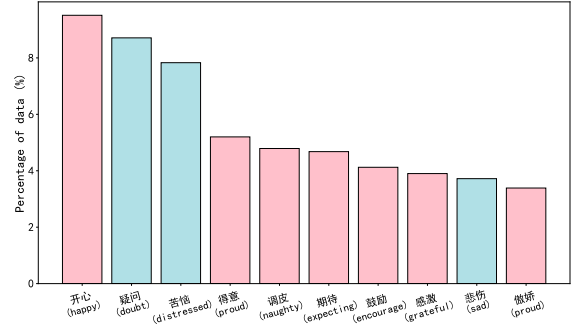


Figure 3: Histogram of top-10 annotated emotions when memes are used. Positive emotions (pink) occur significantly more often than negative emotions (blue).

Considering the limitation of text response evaluation metrics, we selected the top-3 finalists based on the metric score to be manually evaluated for task #1, following the four aspects as:

- *Correctness*: whether there are grammatical errors in the machine generated text response, such as the lack of subject, verb, and *etc.*
- *Relevance*: whether the generated text response related to the historical content of the conversation.
- *Fluency*: whether the generated response is natural and smooth, in line with persons' conversation habits
- *Informativeness*: whether the generated text response contains sufficient information. General replies such as ok and ha ha, are considered to be missing valid information

Besides, we also required annotators to give an *overall score* based on the above four aspects. All of the scores are ranged from 1 to 5 with integers. The annotated data is randomly chosen from the submitted entries of each team, 2000 history-answer pairs for easy and hard test, respectively. The overall score was then aggregated over the entire test sets and we used the average value to determine the official ranking of the dialogue systems in the challenge track.

Results and Analysis

Generally, we received 22 entries in total submitted from 5 participating teams, setting a new state-of-the-art in all three subtasks. To preserve anonymity, the teams were identified by numbers from 1 to 5, while our baseline (Fei et al. 2021) was listed as team 0. Table 3 presents the evaluation results of the best entry from each team in the automatic metrics for different tasks. The full scores with all the submitted entries and metric evaluation scripts are available on the track repository². Most entries achieved comparable performance and outperformed the baseline in all three tasks.

For Task #1: Text Response Modeling, Team 1 performs best in the automatic evaluation outperforming the other

²<https://github.com/lizekang/DSTC10-MOD>

Team	Task #1				Task #2			Task #3			
	B-2	B-4	D-1	D-2	R ₁₀ @1	R ₁₀ @3	R ₁₀ @5	MAP	A@1	A@3	A@5
<i>Easy test</i>											
0	3.35	1.31	1.72	18.2	31.2	54.8	72.1	49.2	53.7	70.1	73.6
1*	5.08	4.25	1.90	26.7	34.2	59.6	76.0	52.3	-	-	-
2*	3.78	1.89	2.20	20.2	34.4	60.4	76.5	52.3	58.3	74.3	78.9
3*	3.57	1.32	1.93	21.5	56.8	84.7	94.4	72.0	62.3	83.4	89.5
4	3.70	1.65	2.00	19.5	33.5	58.0	75.3	50.3	54.5	70.8	74.5
5	3.54	1.40	1.85	20.3	34.0	56.8	74.5	51.0	57.3	72.0	77.6
<i>Hard test</i>											
0	3.15	1.22	1.05	15.8	25.5	49.2	64.0	40.5	25.6	36.0	45.5
1*	5.04	3.65	1.10	20.0	26.8	50.6	65.5	42.3	-	-	-
2*	4.03	1.65	1.36	16.6	27.9	50.8	66.7	45.1	29.7	40.6	49.9
3*	3.65	1.30	1.17	17.7	42.0	69.7	80.9	58.8	27.3	39.2	47.5
4	3.52	1.35	1.00	16.8	27.5	51.0	67.6	50.3	26.5	36.8	46.5
5	3.30	1.26	1.23	15.6	25.7	49.5	63.8	40.9	27.0	37.6	47.2

Table 3: Evaluation results of the best entry. Bold denotes the best results in each column and * indicates the finalists.

Task	Evaluation Metrics
Task #1	BLEU, Dist Human Evaluation
Task #2	Recall _n @k, MAP
Task #3	Accuracy@k

Table 4: Evaluation metrics of the best entry from each team for the Track 1 tasks.

Team	Cor	Rel	Flu	Info	Overall Score
<i>Easy test</i>					
1	3.82	3.88	3.65	3.15	3.74
2	3.68	3.86	3.62	3.23	3.60
3	3.75	3.77	3.67	3.35	3.69
<i>Hard test</i>					
1	3.80	3.75	3.66	3.10	3.68
2	3.72	3.76	3.58	3.18	3.65
3	3.76	3.80	3.70	3.32	3.72

Table 5: Human evaluation results for the Track 1 task #1.

teams by a large margin, with the ensemble several large-scale pre-training models. Team 2 and Team 3 both utilize the PLATO-2 (Bao et al. 2020) for response generation. Table 5 presents the human evaluation results of the task #1 participating teams. We can find that team 1 wins the text response modeling task in easy test set while team 3 achieves the first in the hard test set. Team 1 focuses more on correctness and relevance while Team 3 gains the highest scores in fluency and informativeness. The big gap in automatic evaluation and relatively small gap in human evaluation between teams show that the automatic metrics are not reliable in some degree for the open-domain dialogue. We need to build more reliable automatic evaluation metrics for the task.

For Task #2: Meme Retrieval, Team 3 achieves over 90% Recall₁₀@5 in the easy test set, and also the highest scores in the hard test set. They treat the meme retrieval task as a matching problem and employ the cross-encoder architecture for relevance estimation using negative sampling. Team 2 and Team 3 both choose to utilize the titles of memes instead of the memes themselves, as there is no suitable pre-trained model for meme feature extraction. However, in the real scenario, many memes don’t have accurate titles or one title corresponds to many memes. Memes themselves have the specific meanings. The big gap of performance between easy test and hard test also reveals that the generalization

ability is limited for meme retrieval. So maybe pre-training on a large-scale memes dataset can improve the generalization ability using visual features.

For Task #3: Meme Emotion Classification, Team 3 achieves the highest score of 89.5% in the easy test and 49.9% in the hard test. Particular, they devise an auxiliary method called Emotion-Enhanced Masked LM to improve the ability of meme emotion recognition. Meantime, Team 2 integrated historical memes and constructed good-quality candidate sets to reduce the difficulty of model learning and advance the multimodal content understanding. There is also a big gap between the easy test and hard test. We think the reason is that the emotion is closely related to the visual features which is under explored in the systems. In the easy test, Team 3 achieves a high performance as the titles of memes are all seen before. But in the hard test, the memes are never seen before so that it’s hard to recognize the emotion only based on the titles.

Conclusion

In this paper, we describe the task definition, provided datasets, and evaluation set-up for DSTC10-MOD tracks. We also summarize the results of the submitted system, which highlight overall trends of the state-of-the-art techniques for multi-model dialogue. First, the top systems are all built with transformer-based end-to-end learning and follow the pre-training and fine-tuning paradigm. Large-scale pre-training technology has become the mainstream. Second, incorporation of extra data for contrastive learning can effectively improve the robustness and generalization of the model, *e.g.*, its ability for applying unseen internet memes. Third, well-designed self-supervised tasks can boost the multi-modal information fusion and understanding of the system, *e.g.*, Emotion-Enhanced Masked Language Modeling task in team 3 obtains a notable performance gain in meme retrieval. Although there is a lot of advancement compared with the baseline, we believe that the MOD task is worth further exploring and can benefit the modeling of multi-modal open-domain dialogue intelligence in the future, especially in how to better exploit the visual features of memes.

References

- AlAmri, H.; Cartillier, V.; Das, A.; Wang, J.; Cherian, A.; Essa, I.; Batra, D.; Marks, T. K.; Hori, C.; Anderson, P.; Lee, S.; and Parikh, D. 2019. Audio Visual Scene-Aware Dialog. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 7558–7567. Computer Vision Foundation / IEEE.
- Bao, S.; He, H.; Wang, F.; Wu, H.; Wang, H.; Wu, W.; Guo, Z.; Liu, Z.; and Xu, X. 2020. Plato-2: Towards building an open-domain chatbot via curriculum learning. *arXiv preprint arXiv:2006.16779*.
- Beskow, D. M.; Kumar, S.; and Carley, K. M. 2020. The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning. *Inf. Process. Manag.*, 57(2): 102170.
- Chen, C. 2020. Research on Sticker Cognition for Elderly People Using Instant Messaging. In Rau, P. P., ed., *Cross-Cultural Design. User Experience of Products, Services, and Intelligent Environments - 12th International Conference, CCD 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19-24, 2020, Proceedings, Part I*, volume 12192 of *Lecture Notes in Computer Science*, 16–27. Springer.
- Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M. F.; Parikh, D.; and Batra, D. 2017. Visual Dialog. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 1080–1089. IEEE Computer Society.
- Fei, Z.; Li, Z.; Zhang, J.; Feng, Y.; and Zhou, J. 2021. Towards Expressive Communication with Internet Memes: A New Multimodal Conversation Dataset and Benchmark. *arXiv preprint arXiv:2109.01839*.
- Posey, C.; Lowry, P. B.; Roberts, T. L.; and Ellis, T. S. 2010. Proposing the online community self-disclosure model: the case of working professionals in France and the U.K. who use online communities. *Eur. J. Inf. Syst.*, 19(2): 181–195.
- Wang, Y.; Ke, P.; Zheng, Y.; Huang, K.; Jiang, Y.; Zhu, X.; and Huang, M. 2020. A Large-Scale Chinese Short-Text Conversation Dataset. In Zhu, X.; Zhang, M.; Hong, Y.; and He, R., eds., *Natural Language Processing and Chinese Computing - 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14-18, 2020, Proceedings, Part I*, volume 12430 of *Lecture Notes in Computer Science*, 91–103. Springer.
- Wang, Y.; Li, Y.; Gui, X.; Kou, Y.; and Liu, F. 2019. Culturally-Embedded Visual Literacy: A Study of Impression Management via Emoticon, Emoji, Sticker, and Meme on Social Media in China. *Proc. ACM Hum. Comput. Interact.*, 3(CSCW): 68:1–68:24.