Overview of the NTCIR-14 Short Text Generation Subtask: Emotion Generation Challenge

Yaoqin Zhang[1] and Minlie Huang[1]

Institute for Artificial Intelligence,
State Key Lab of Intelligent Technology and Systems,
Beijing National Research Center for Information Science and Technology,
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
zhangyq17@mails.tsinghua.edu.cn aihuang@tsinghua.edu.cn

Abstract This paper describes an overview of the Emotion Generation subtask at NTCIR-14. The goal of the emotion generation subtask is to investigate how well a chatting machine can express feelings by generating a textual response to an input post. The task is defined as follows: given a post and a pre-specified emotion class of the generated response, the task is to generate a response that is appropriate in both topic and emotion. This challenge has attracted more 40 teams registered, and 11 teams finally submitted results. In this overview paper, we reported the details of this challenge, including task definition, data preparation, annotation schema, submission statistics, and evaluation results.

1 Introduction

During the past years, there has been a developing trend in AI research to enhance Human-Computer Interaction by humanizing machines. However, to create a robot capable of acting and talking with a user at the human level requires the robot to understand human cognitive behaviors, while one of the most important human behaviors is understanding and expressing emotions and affects. As a vital part of human intelligence, emotional intelligence is defined as the ability to perceive, integrate, understand, and regulate emotions.

In recent years, deep learning approaches have advanced dialogue/conversation generation significantly. Thanks to the success of neural generative networks [1, 13], these models were soon applied to conversation generation [14], including the neural responding machine [10], the hierarchical recurrent encoder-decoder neural network [9], and many others [12]. Existing works mainly focus on improving the content quality of generated responses by employing superior decoding strategies [4, 6, 11]. Also there are some works on applying stronger control in generation via label or word type variables [3, 15], or by embedding knowledge into dialog generation [18, 21]. Other attempts to improve content quality include considering additional topic words [7,16], topic categories [17], and persona information [5, 8].

Though a variety of models have been proposed for large-scale conversation generation, it is still quite challenging (and yet to be addressed) to generate emotional responses. Recently, Zhou et al. [20] proposed an emotional chatting machine that is able to generate responses that are appropriate not only in content but also in emotion expression. Inspired

by this work, we defined the challenge task in this paper accordingly, but we reformulated the task with a new dataset. This is the second time we host emotion generation challenge [2].

In this challenge, participants are expected to generate Chinese responses that are both fluent in content and adequate in emotion, which is quite critical for building an empathic chatting machine. For example, if a user says "I had a terrible day.", the chatting should respond with "It will be a great day tomorrow." to express *comfort*, but it could also be "I'm sorry to hear that, what happened?" to express *sadness*.

2 Task Definition

The task is defined as follows: Given a Chinese post $X = (x_1, x_2, \cdots, x_n)$, and a pre-specified emotion category of the response to be generated, the goal is to generate a response $Y = (y_1, y_2, \cdots, y_m)$ that is relevant to post and also coherent with the emotion category. The emotion classes are in {*Anger, Disgust, Happiness, Like, Sadness, Other*}. Exemplar responses are shown as below:

Table 1. Conversations with/without considering emotion [20]

User: Worst day ever. I arrived late because of the traffic.
Basic Seq2Seq: You were late.
ECM (Like): I am always here to support you.
ECM (Happy): Keep smiling! Things will get better.
ECM (Sad): It's depressing.
ECM (Disgust): Sometimes life just sucks.
ECM (Angry): The traffic is too bad!

3 Dataset Description

The dataset is built from Weibo posts and replies/comments. More than 1 million Weibo post-response pairs are provided to participants for training their models. The test dataset consists of 200 manually filtered posts. For each post and response, we used a classifier to automatically label its emotion class. Participating systems should generate a response for each emotion class. Note that participants should generate responses for all emotion classes for each post.

During the construction of the dataset, we trained a bidirectional LSTM model classifier to automatically label the posts and responses with emotion. The classifier was trained on the data from the NLPCC 2013/2014 Emotion Classification Challenge. The accuracy of our classifier for six-way classification is about 62%, for more details please refer to [20].

With the help of the emotion classifier, we selected those responses that have a small value of classification entropy, defined as follows:

$$CE = -\frac{1}{\log K} \sum_{k=1}^{K} p_k \log p_k \tag{1}$$

where p_k is the probability of class k given by the emotion classifier, and K is the total number of classes. Note that $0 \leq CE \leq 1$, and **0** indicates the least uncertainty of the prediction. In this way, we can select those pairs whose responses have obvious emotional expressions.

Due to the heavy annotation load, we provided 200 manually filtered posts, 40 for each emotion class except for *other*. These 200 posts were selected according to the following guideline:

— The post should not be context-dependent. In other words, understanding of the post does not require any other context or background knowledge.

— The post should not include rare words or English words.

For each post, participants should generate one response for each emotion class, except for *Other*.

It is worth noting that, the emotion label of these data is noisy. Participants are encouraged to implement their own emotion classifiers and train with their own data, as long as all the details are reported and all resources will be accessible to the community to let other researchers reproduce their results. Please notice that no external resource data can be used to train the response generation models.

4 Annotation Schema

The submitted post-response pairs are evaluated by the following metrics:
Emotion Consistency: whether the emotion class of a generated response is the same as the pre-specified class.
Coherence: whether the response is appropriate in terms of both logically coherent and topic relevant content.
Fluency: whether the response is fluent in grammar and acceptable as a natural language response.

Our labeling procedure is shown by the following pseudo code [19]:

Algorithm 1 Labeling procedure

```
if (Coherence and Fluency) then
    if (Emotion Consistency) then
        LABEL 2 ## Score 2 for perfect responses
    else
        LABEL 1 ## Score 1 for coherent and fluent responses
    end if
else
    LABEL 0 ## Score 0 for others
end if
```

Following are some annotation examples in Table 2.

Table 2. Annotation examples.

Post	Men that can cook are so cool! 会做饭的男人最帅了！	Emotion class	Coherence/Fluency	Emotion Consistency	Label
Response 1	Man that can cook are sure cool! 会做饭的男人是帅啊！	Like	Yes	Yes	2
Response 2	lol, I agree. LOL，我同意。	Like	Yes	No	1
Response 3	Which movie is this from? 这是哪部电影中的呀?	Disgust	No	Yes	0
Response 4	Same to you. 你也是。	Like	No	No	0
Response 5	I love you will 我爱你会	Disgust	No	Yes	0
Response 6	This is same way dotcorine of ways. 这是同道主义的道	Disgust	No	No	0

Particularly, for those repeatedly occurred contents in a response, if a subsequence of content occurs repeatedly for no more than 3 times, it will be judge as fluent, otherwise not fluent. Some examples are shown in Table 3.

Table 3. The examples of the fluency judgement on responses with repetitive words.

Response	Fluency
Sad, sad, sad. 难过，难过，难过。	Yes
So cute, so cute, so cute 好可爱，好可爱，好可爱	yes
Yes yes yes yes yes yes yes. 是是是是是是是	No
None of my business none of my business none of my business none of my business 不关我事关我事关我事关我事	No
China's education education education education 中国的教育教育教育教育	No

5 Submission Statistics

We received 21 submissions from 11 teams before the deadline of submission, 16 of them are in valid format, 1 matches format requirement but does not generate meaningful response. Each submission contain 1000 responses, 5 responses for each post.

6 Evaluation Results

The submitted results from all teams are aggregated together. After de-duplication, we obtained 15,263 post-response pairs. Then, these pairs are randomly shuffled with the

submission identifier for each pair recorded. We resorted to Baidu Data Crowdsourcing Service for manual evaluation. Each pair is annotated by three curators who were trained with our annotation schema and illustrating examples. The annotation statistics are shown in Table 4.

Table 4. The statistics of annotation agreement.

	All three agree	Two agree	All different
Number of pairs	9,527	5,150	626
Percentage	62%	34%	4%

We can see that 626 of all the pairs were assigned to different labels by all three annotators. For these pairs, we considered the label as 0, since the quality seems to be not reliable. Other pairs' labels are decided by the majority voting rule. 96% of all pairs receive the same label by at least 2 annotators, which is fairly good agreement.

For each submission run, we computed overall score and average score. The formulas are listed as below:

$$OverallScore = \sum_{i=0}^{2} i * num_i \tag{2}$$

$$AverageScore = \frac{1}{N_t} \sum_{i=0}^{2} i * num_i \tag{3}$$

where Num_i is the number of pairs which have a label of i in each submission run, and N_t is the total number of submitted pairs for each run.

6.1 Overall Results

The overall results are presented in Table 5. We can see that the best result has a score of 0.953 (from RUCIR). There are two runs that score about 0.81∼0.82 and other two runs scoring 0.72∼0.74. For other runs, the results are apparently much lower than these top-performing results.

We also observed extremely low scores in some submissions. For instance, in WUST_2, the low score is because there are only 5 unique responses in all the responses and the responses are irrelevant. In IMTKU_2, the reason for the low score may be due to their method: the team searched candidate responses in a small set of candidates.

Table 5. The result of the overall score and average score.

Team Name	Label 0	Label 1	Label 2	Total	Overall score	Average score
1194_1	581	320	99	1000	518	0.518
1194_2	831	109	60	1000	229	0.229
AINTPU_1	716	200	84	1000	368	0.368
CKIP_1	845	29	126	1000	281	0.281
CKIP_2	840	28	132	1000	292	0.292
IMTKU_1	580	248	172	1000	592	0.592
IMTKU_2	954	32	14	1000	60	0.06
RUCIR_1	392	263	345	1000	953	0.953
RUCIR_2	460	342	198	1000	738	0.738
TMUNLP_1	777	126	97	1000	320	0.32
TUA1_1	443	293	264	1000	821	0.821
TUA1_2	454	278	268	1000	814	0.814
WUST_1	601	211	188	1000	587	0.587
WUST_2	999	0	1	1000	2	0.002
TKUIM_2	507	260	233	1000	726	0.726
CYIII_1	617	267	116	1000	499	0.499

6.2 Emotion-specific results

We also computed the scores for each emotion category, in order to investigate how the models perform on different emotion classes. The result is listed in the table below.

We can see that average scores between different emotions are not that remarkable. However, this may be because the submissions focus more on how to respond to a post but not to control emotion. If we focused on the submissions with label 2, which means correct emotion consistency, we can see that *Anger* and *Disgust* have much lower scores. This may be due to the lack of sufficient data with corresponding emotion classes in the corpus. This observation is consistent to [20] which reports worse generation performance on the minor emotion categories.

7 Models from Submission Teams

In this section, we will make a brief overview on the model summaries from the technology perspective. Those models with higher scores does not consider much about emotion consistency, this may be because achieving both emotion consistency and coherence is much harder than consistency only. Table 11 shows the model summaries and overall scores from each submission.

1194_1 and 1194_2 employs the same seq2seq model on two different datasets. 1194_1 uses the whole dataset while 1194_2 uses only post-response pairs of the same emotion class. Participants consider _2 may be less noisy on emotion label and can reach better

Table 6. The result on the emotion category of *Like*.

Team Name	Label 0	Label 1	Label 2	Total	Overall score	Average score
1194_1	119	42	39	200	120	0.6
1194_2	170	11	19	200	49	0.245
AINTPU_1	153	16	31	200	78	0.39
CKIP_1	164	2	34	200	70	0.35
CKIP_2	171	5	24	200	53	0.265
IMTKU_1	119	35	46	200	127	0.635
IMTKU_2	194	4	2	200	8	0.04
RUCIR_1	88	36	76	200	188	0.94
RUCIR_2	96	44	60	200	164	0.82
TMUNLP_1	164	9	27	200	63	0.315
TUA1_1	121	11	68	200	147	0.735
TUA1_2	109	24	67	200	158	0.79
WUST_1	117	36	47	200	130	0.65
WUST_2	199	0	1	200	2	0.01
TKUIM_2	90	56	54	200	164	0.82
CYIII_1	138	33	29	200	91	0.455

Table 7. The result on the emotion category of *Sad*.

Team Name	Label 0	Label 1	Label 2	Total	Overall score	Average score
1194_1	112	73	15	200	103	0.515
1194_2	158	22	20	200	62	0.31
AINTPU_1	135	60	5	200	70	0.35
CKIP_1	165	4	31	200	66	0.33
CKIP_2	159	2	39	200	80	0.4
IMTKU_1	116	48	36	200	120	0.6
IMTKU_2	189	5	6	200	17	0.085
RUCIR_1	72	48	80	200	208	1.04
RUCIR_2	83	57	60	200	177	0.885
TMUNLP_1	163	26	11	200	48	0.24
TUA1_1	84	31	85	200	201	1.005
TUA1_2	92	40	68	200	176	0.88
WUST_1	124	31	45	200	121	0.605
WUST_2	200	0	0	200	0	0.0
TKUIM_2	115	40	45	200	130	0.65
CYIII_1	105	60	35	200	130	0.65

Table 8. The result on the emotion category of *Disgust*.

Team Name	Label 0	Label 1	Label 2	Total	Overall score	Average score
1194_1	109	83	8	200	99	0.495
1194_2	167	30	3	200	36	0.18
AINTPU_1	140	56	4	200	64	0.32
CKIP_1	183	2	15	200	32	0.16
CKIP_2	179	5	16	200	37	0.185
IMTKU_1	117	69	14	200	97	0.485
IMTKU_2	193	7	0	200	7	0.035
RUCIR_1	71	76	53	200	182	0.91
RUCIR_2	90	96	14	200	124	0.62
TMUNLP_1	158	42	0	200	42	0.21
TUA1_1	82	105	13	200	131	0.655
TUA1_2	92	82	26	200	134	0.67
WUST_1	111	69	20	200	109	0.545
WUST_2	200	0	0	200	0	0.0
TKUIM_2	89	96	15	200	126	0.63
CYIII_1	120	51	29	200	109	0.545

Table 9. The result on the emotion category of *Anger*.

Team Name	Label 0	Label 1	Label 2	Total	Overall score	Average score
1194_1	114	81	5	200	91	0.455
1194_2	158	42	0	200	42	0.21
AINTPU_1	150	45	5	200	55	0.275
CKIP_1	164	12	24	200	60	0.3
CKIP_2	159	8	33	200	74	0.37
IMTKU_1	124	64	12	200	88	0.44
IMTKU_2	189	11	0	200	11	0.055
RUCIR_1	88	63	49	200	161	0.805
RUCIR_2	98	91	11	200	113	0.565
TMUNLP_1	154	42	4	200	50	0.25
TUA1_1	85	110	5	200	120	0.6
TUA1_2	85	107	8	200	123	0.615
WUST_1	137	48	15	200	78	0.39
WUST_2	200	0	0	200	0	0.0
TKUIM_2	112	45	43	200	131	0.655
CYIII_1	122	71	7	200	85	0.425

Table 10. The result on the emotion category of *Happy*.

Team Name	Label 0	Label 1	Label 2	Total	Overall score	Average score
1194_1	127	41	32	200	105	0.525
1194_2	178	4	18	200	40	0.2
AINTPU_1	138	23	39	200	101	0.505
CKIP_1	169	9	22	200	53	0.265
CKIP_2	172	8	20	200	48	0.24
IMTKU_1	104	32	64	200	160	0.8
IMTKU_2	189	5	6	200	17	0.085
RUCIR_1	73	40	87	200	214	1.07
RUCIR_2	93	54	53	200	160	0.8
TMUNLP_1	138	7	55	200	117	0.585
TUA1_1	71	36	93	200	222	1.11
TUA1_2	76	25	99	200	223	1.115
WUST_1	112	27	61	200	149	0.745
WUST_2	200	0	0	200	0	0.0
TKUI_1	101	23	76	200	175	0.875
CYIII_1	132	52	16	200	84	0.420

quality while the size reduced (by a factor of 4). The result shows that smaller training datasets lead to lower fluency in output.

RUCIR_1 is the only submission with a hybrid strategy that combines rule-based and generation models, and RUCIR_2 is the generation part of RUCIR_1. The team implemented a complicated generation model, with seq2seq model and copy mechanism. The generation model itself obtained a score of 738. After including the rule-based module, its score increases by over 200 and reaches 953. This indicates that the rule-based model can effectively increase the quality of output under this circumstance.

TUA1_1 and TUA1_2 obtained a score of 800 using generation models. These two submissions concatenate emotion category information with character-level seq2seq models. Comparing to TUA1_1, TUA1_2 uses Weibo posts of the same emotion class as additional input. Although the overall score of TUA1_2 is not significantly higher than that of TUA1_1, the responses generated by TUA1_2 are more informative and relevant to the original post, according to the cases provided by the team.

IMTKU_2 used a generation model. The model generated 700 responses for each emotion class, and used a reranking model to choose the most related response. Searching responses in such a small set lead to low relevance between post and response.

Table 11. Model summaries of each team

Team Name	Method	Model Structure	Special Feature	Overall score
1194_1	Generation	Seq2Seq (Bi-LSTM + Attention)	Emotion Context Vector Emotion SOS Tag	518
1194_2	Generation	Seq2Seq (Bi-LSTM + Attention)	Emotion Context Vector Emotion SOS Tag	229
AINTPU_1	Retrieval	Two-step Ranking	TF-IDF Cosine Similarity	368
CKIP_1	Generation	Seq2Seq (Bi-LSTM + Attention) Forward Only Decoder	Emotion Keywords	281
CKIP_2	Generation	Seq2Seq (Bi-LSTM + Attention) Forward and Backward Decoder	Emotion Keywords	292
IMTKU_1	Retrieval	Three-step Ranking	Solr Matching Emotion Matching Word2Vec Similarity	592
IMTKU_2	Generation	Seq2Seq (LSTM) Reranking	Emotion Matching Word2Vec Similarity	60
RUCIR_1	Hybrid (Rule-based + Generation)	Rule-Based Seq2Seq (GRU + Attention) Copy Mechanism Reranking	Emotion Embedding	953
RUCIR_2	Generation	Seq2Seq (GRU + Attention) Copy Mechanism	Emotion Embedding	738
TMUNLP_1	Generation	Seq2Seq (Bi-LSTM) Refinement	Emotion Vector Emotion Words	320
TUA1_1	Generation	Seq2Seq (Bi-LSTM)	Emotion Embedding	821
TUA1_2	Generation	Seq2Seq (Bi-LSTM)	Emotion Embedding Post of Same Emotion	814
WUST_1	Retrieval	Two-step Ranking	Emotion Class TF-IDF + VSM	587
TKUIM_2	Generation	Seq2Seq (Bi-LSTM + Attention) Multi Generator Reranking		726

8 Summary

In this paper, we presented the task definition, datasets, evaluation metrics, and results for the emotional conversation generation challenge. This is the second challenge of letting a chatting machine to express emotion via textual output in the setting of large-scale conversation generation. Comparing to the results of the last challenge, we find that there is still a long way to produce satisfactory results.

— The proportion of zero scored results is large, demonstrating that most replies are not appropriate in content or in emotion.

— None of the submission runs reaches at an average score of 1.0, and this shows that the current performance of all submissions is still far from satisfactory.

— We required participants to generate responses for all emotion classes. However, for some specific posts, it can be much harder to generate responses for some emotion classes (for instance, *Angry* and *Disgust*) than others.

— Some of the top performing submissions adopt retrieval-based results, implying that generation-based models still have much room to improve.

9 Acknowledgement

We thank the STC-3 participants and the NTCIR chairs for making this task happen. This work was supported by the National Key R&D Program of China (Grant No. 2018YFC0830200), and partly by the National Science Foundation of China (Grant No.61876096/61332007).

We would like to thank Prof. Xiaoyan Zhu for her unreserved support.

Reference

1. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
2. Minlie Huang, Zuoxian Ye, and Hao Zhou. Overview of the NLPCC 2017 shared task: Emotion generation challenge. In Xuanjing Huang, Jing Jiang, Dongyan Zhao, Yansong Feng, and Yu Hong, editors, *Natural Language Processing and Chinese Computing - 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8-12, 2017, Proceedings*, volume 10619 of *Lecture Notes in Computer Science*, pages 926–936. Springer, 2017.
3. Pei Ke, Jian Guan, Minlie Huang, and Xiaoyan Zhu. Generating informative responses with controlled sentence function. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1499–1508. Association for Computational Linguistics, 2018.
4. Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119. The Association for Computational Linguistics, 2016.
5. Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and William B. Dolan. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016.
6. Jiwei Li, Will Monroe, and Dan Jurafsky. A simple, fast diverse decoding algorithm for neural generation. *CoRR*, abs/1611.08562, 2016.
7. Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 3349–3358. ACL, 2016.

8. Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. Assigning personality/profile to a chatting machine for coherent conversation generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 4279–4285. ijcai.org, 2018.
9. Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. Hierarchical neural network generative models for movie dialogues. *CoRR*, abs/1507.04808, 2015.
10. Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1577–1586. The Association for Computer Linguistics, 2015.
11. Louis Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. Generating long and diverse responses with neural conversation models. *CoRR*, abs/1701.03185, 2017.
12. Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 196–205. The Association for Computational Linguistics, 2015.
13. Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112, 2014.
14. Oriol Vinyals and Quoc V. Le. A neural conversational model. *CoRR*, abs/1506.05869, 2015.
15. Yansen Wang, Chenyi Liu, Minlie Huang, and Liqiang Nie. Learning to ask questions in open-domain conversational systems with typed decoders. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2193–2203. Association for Computational Linguistics, 2018.
16. Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. Topic aware neural response generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 3351–3357. AAAI Press, 2017.
17. Kun Xiong, Anqi Cui, Zefeng Zhang, and Ming Li. Neural contextual conversation learning with labeled question-answering pairs. *CoRR*, abs/1607.05809, 2016.
18. Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4970–4977. AAAI Press, 2018.
19. Zhaohao Zeng, Sosuke Kato, and Tetsuya Sakai. Overview of the NTCIR-14 short text conversation task: Dialogue quality and nugget detection subtasks. In *NTCIR-14, 2019.*
20. Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 730–739. AAAI Press, 2018.
21. Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 4623–4629. ijcai.org, 2018.