Explicit Use of Topicality in Dialogue Response Generation

Takumi Yoshikoshi Hayato Atarashi Takashi Kodama Sadao Kurohashi
Kyoto University, Japan
{takumiyoshiko, atarashi, kodama}@nlp.ist.i.kyoto-u.ac.jp
kuro@i.kyoto-u.ac.jp

Abstract

The current chat dialogue systems implicitly consider the topic given the context, but not explicitly. As a result, these systems often generate inconsistent responses with the topic of the moment. In this study, we propose a dialogue system that responds appropriately following the topic by selecting the entity with the highest "topicality." In topicality estimation, the model is trained through self-supervised learning that regards entities appearing in both context and response as the topic entities. In response generation, the model is trained to generate topic-relevant responses based on the estimated topicality. Experimental results show that our proposed system can follow the topic more than the existing dialogue system that considers only the context.

1 Introduction

In recent years, end-to-end chat dialogue systems have been developed remarkably, making it possible to generate rich and flexible responses. However, such current chat dialogue systems only implicitly consider the topic given the context as it is, but do not explicitly consider it (Adiwardana et al., 2020; Roller et al., 2021). As a result, these systems often generate inconsistent responses with the topic (Sugiyama et al., 2021).

In this study, we propose a dialogue system that selects the next dialogue topic and responds using the selected topic explicitly. Our system selects the next topic entity based on **topicality** (Givón, 1983). Here, we define the entity as a noun or compound nouns, and topicality as the degree of speaker awareness directed toward each entity in the dialogue context. In addition, we call the entity with the highest topicality in the context **topic entity**. In the response generation part, our system generates responses based on the estimated topic entity as well as the context.

We propose the **Two-Stage** model, which learns topicality estimation and response generation in

two stages, and the **End-to-End** model, which learns in the end-to-end method.

Due to the lack of dialogue corpus with the topic annotated, we use a self-supervised learning method to train our proposed models. Specifically, we extract triples of <*context*, *response*, *labeled topic entity candidates*> from the unannotated dialogue corpus.

For labeling, we regard topic entity candidates (= entities in the context) in the response as topic entities and assign labels to them. This procedure assumes that the entity in the response can be considered the topic entity. Furthermore, zero anaphora resolution is applied to restore those omitted words when assigning labels because word omission is pervasive in actual dialogue (especially in Japanese).

The automatic and human evaluation results show that our proposed system can follow the topic more than the existing dialogue system that considers only the context.

2 Related Work

2.1 Dialogue Systems that Consider only Context

Existing dialogue systems that consider only the dialogue context sometimes generate dull responses for elevating a naturalness of response (Vinyals and Le, 2015; Shang et al., 2015). To solve this issue, dialogue systems based on the Transformer (Vaswani et al., 2017), such as Meena (Adiwardana et al., 2020) and BlenderBot (Roller et al., 2021), have been proposed. These dialogue systems generate diverse and engaging responses with large dialogue data and model parameters.

However, the above dialogue systems, which only consider the context, do not consider the topic explicitly and may generate inconsistent responses with the topic (Sugiyama et al., 2021). In this study, we construct dialogue systems that explicitly con-

222

	# of dialogues	# of utterances	# of triples w/o ZAR	# of triples w/ ZAR
Twitter corpus	250M	670M	-	-
JPersonaChat	5,000	61,794	23,217 / 147,166	27,809 / 176,122
JEmpatheticDialogues	20,000	80,000	16,269 / 64,999	31,466 / 114,888
KUHCC	5,114	86,192	21,503 / 120,770	35,467 / 199,094

Table 1: Statistics for each dialogue corpus. ZAR stands for Zero Anaphora Resolution. The left sides of columns "w/o ZAR" and "w/o ZAR" show the number of triples with the positive label, and the right side shows the number of triples with the negative label.

sider topicality to generate responses following the topic.

2.2 Dialogue Systems that Explicitly Consider the Topic

There are two purposes for explicitly considering the topic: generating informative responses and generating responses following the topic.

To generate informative responses, Xing et al. (2017) proposed a dialogue system considering the topic explicitly. This system predicts topic words that are highly relevant to words in the context by a pretrained Twitter LDA model (Zhao et al., 2011) and generates responses based on the predicted topic words. Mou et al. (2016) proposed a system that selects the noun with the highest PMI (Church and Hanks, 1990) against words in the context and generates responses that contains the noun.

To generate responses following the topic, Zhang et al. (2020) attempts to generate topic-relevant responses by learning examples in which entities in the dialogue context continue to appear in the response as the topic. However, in actual dialogues, the topic entity is omitted more frequently than other words (Givón, 1983). The method of Zhang et al. (2020) does not consider this omission problem, but we consider it by restoring the omitted entities.

3 Dataset Construction

We construct the dataset for self-supervised training of the proposed model. The dataset is constructed based on the assumption that entities that appear in both context and response are the topic entities. Considering the pervasiveness of word omission in actual dialogue, the omitted words are restored by zero anaphora resolution.

3.1 Dialogue Corpora

All models used in this study are pre-trained by the Twitter Corpus and then fine-tuned by JPersonaChat (Sugiyama et al., 2021), JEmpatheticDi-

alogues (Sugiyama et al., 2021), and Kyoto University Hobby Chat Corpus (KUHCC). The size of each corpus is shown in Table 1.

JPersonaChat (Sugiyama et al., 2021) is a dialogue corpus between two Japanese speakers with specific personas based on PesonaChat (Zhang et al., 2018). JEmpatheticDialogues (Sugiyama et al., 2021) is a dialogue corpus between two Japanese speakers talking about an event based on diverse emotional expressions referring to EmpatheticDialogues (Rashkin et al., 2019).

In addition to these two existing corpora, we collect a chat dialogue corpus about hobbies, KUHCC, which is collected by crowdsourcing[1]. For dialogue collection, we use existing dialogue collection framework [2]. In this framework, when workers access the specified URL for dialogue collection, pair-matching is performed automatically, and a chat room is created for the workers to interact in real-time. It is challenging to get the workers to chat completely freely, therefore, paired workers are assigned different roles: one is the speaker, and the other is the listener. The speaker talks about their hobbies, and the listener listens while asking questions about the speaker's hobbies.

3.2 Method

We extract triples of <*context*, *response*, *labeled topic entity candidates*> from dialogue corpora by the self-supervised method.

For each context-response pair, up to 8 recently used nouns are extracted from the context and used as topic entity candidates. Personal pronouns and interrogatives are removed, and consecutive nouns in the same clause are extracted together as compound nouns. If an entity appears multiple times in a context, only the last entity is extracted. Juman++ (Tolmachev et al., 2018) and BERTKNP[3]

---

[1] https://crowdsourcing.yahoo.co.jp/
[2] https://github.com/ku-nlp/ChatCollectionFramework
[3] https://github.com/ku-nlp/bertknp

were used in this process.

In order to restore the word omissions in the response, we applied zero anaphora resolution using the Cohesion Analysis model (Ueda et al., 2020). This model was trained with multiple Japanese semantic relation analysis tasks: predicate-argument structure analysis, bridging anaphora resolution, and coreference resolution.

The topic entity candidates that appear in the response, including the restored one, are then assigned the "positive" label, which indicates that the entity is the topic entity. On the contrary, we assign the "negative" label, which indicates that the entity is not the topic entity, to the topic entity candidates that do not appear in the response. If there is more than one positive label in the context, only the last entity in the context is assigned the positive label[4].

3.3 Statistics

The models do not learn topicality estimation[5] in pre-training, hence we only extract context-response pairs from the Twitter Corpus. For fine-tuning, we extract triples from each of the three corpora, using the method described in Section 3.2. Table 1 shows the statistics of the constructed dataset. By restoring omitted words, we can obtain 33,548 more triples for JPersonaChat, 65,086 for JEmpatheticDialogues, and 92,288 for KUHCC.

4 Model

In this section, we describe the Two-Stage model, in which topicality estimation and response generation are learned in two stages (Section 4.1), and the End-to-End model, in which they are learned in the end-to-end method (Section 4.2).

4.1 Two-Stage Model

The Two-Stage model generates responses in two stages: Stage 1 and Stage 2 (Figure 1). In Stage 1 (topicality estimation), the model selects a topic entity from topic entity candidates based on the dialogue context. In Stage 2 (response generation), the model generates the response based on the context and the topic entity selected in Stage 1. Note that the gold topic entity is used during the training phase.

[4]In our preliminary experiments, the method, in which all topic entity candidates in the response are considered to be positive, did not get good results.

[5]The method for learning topicality estimation using Twitter corpus did not yield good results in preliminary experiments.

4.1.1 Model Architecture

We use BERT (Devlin et al., 2019) as the topicality estimation model in Stage 1. The input to the model is a topic entity candidate and the utterances in the context in sequence across [SEP] tokens. All topic entity candidates are input in the same way, and the topic entity candidate with the highest output for each [CLS] token is selected as the topic entity.

We use the encoder-decoder model with BERT as the encoder and Transformer (Vaswani et al., 2017) as the decoder in Stage 2. The encoders encode the context and the topic entity separately and then concatenate them. The parameters of these encoders are shared. In the decoder, we additionally use the rewarding mechanism (Takebayashi et al., 2018) to increase the generation probability of topic entities and attempt to generate responses that reflect topic entities.

4.1.2 Loss Function

In Stage 1, the topicality estimation model is trained by minimizing the cross-entropy loss between the probability distribution of the prediction and the gold label.

In Stage 2, the encoder-decoder model is trained by minimizing the following loss function $\mathcal{L}_{nll}$:

$$\mathcal{L}_{nll} = -\sum_{t=1}^{T'} \log p(y_t|y_{<t}, \mathbf{x}, \mathbf{e}), \qquad (1)$$

where T' is the length of the target response, $y_{<t}$ is previously generated sequence, $\mathbf{x}$ is the context, and $\mathbf{e}$ is the topic entity.

4.2 End-to-End Model

The End-to-End model learns topicality estimation and response generation simultaneously (Figure 1). The topicality is estimated using the hidden states of topic entity candidates extracted from the encoded context. The response is generated based on the topic vector calculated based on topicality and the context vector.

4.2.1 Model Architecture

We use BERT as the encoder. The input to the encoder is the contexts split by [SEP] token. We additionally insert a special token [NO_ENTITY] at the beginning of the contexts. The encoder outputs the context vectors: $\mathbf{x} = [\mathbf{x}_1, ..., \mathbf{x}_M]^T \in \mathbf{R}^{M \times d}$ (M is the length of the context).

We then obtain the entity vectors: $\mathbf{e} = [\mathbf{e}_1, ..., \mathbf{e}_{N+1}]^T \in \mathbf{R}^{(N+1) \times d}$ (N is the number
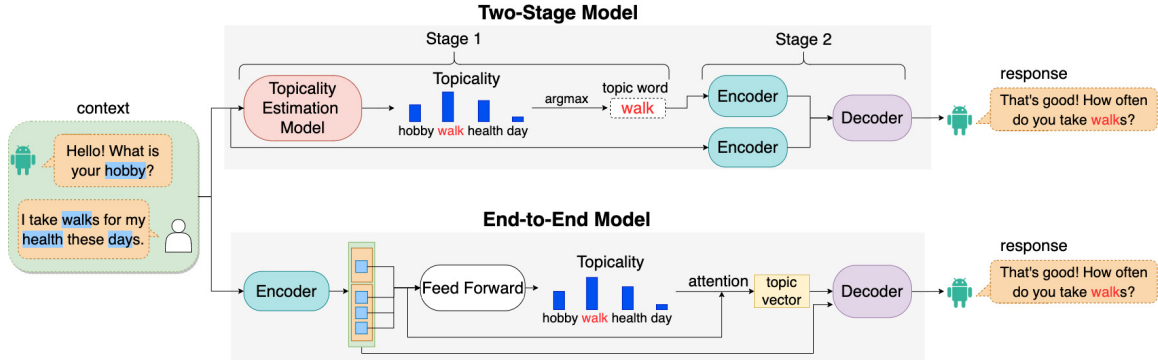
Figure 1: Overview of our proposed models. The Two-Stage model is trained in two stages, Stage 1 (topicality estimation) and Stage 2 (response generation). The End-to-End model learns topicality estimation and response generation simultaneously.

of topic entity candidates) by extracting and concatenating the corresponding vectors of each topic entity candidate in the the context vectors. In the case of a multi-token entity, we take the average of context vectors of the corresponding tokens. Given $\mathbf{e}$ as input, topicality is formulated as follows:

$$P_{topic}(\mathbf{e}) = softmax(\mathbf{e}\mathbf{W}_{topic}) \in \mathbb{R}^{(N+1)\times 1}, \quad (2)$$

where $\mathbf{W}_{topic} \in \mathbb{R}^{d\times 1}$ is a learnable linear layer. The topic vector is calculated using dot-product attention between $P_{topic}(\mathbf{e})$ and $\mathbf{e}$.

The input to the decoder is the concatenated vector of $\mathbf{x}$ and $\mathbf{v}_{topic}$. Similar to the Two-Stage model, the rewarding mechanism is incorporated in the decoder process.

4.2.2 Loss Function

We combine the negative log-likelihood loss for response generation $\mathcal{L}_{nll}$ and the cross-entropy loss for topicality estimation $\mathcal{L}_{topic}$ modulated by a weight by α, which is the hyperparameter. The overall loss function $\mathcal{L}$ is:

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{nll} + \alpha\mathcal{L}_{topic} \quad (3)$$

Note that $\mathcal{L}_{nll}$ is the same as in equation (1), and in the loss function for the topicality estimation, the parameters are not updated if the corresponding label is negative (= the topic entity candidate is not in the response.)

5 Experiment

5.1 Experimental Settings

We use the Japanese pre-trained BERT Large model with whole word maskifng[6] as the encoder.

For the decoder, we use a 12-layer Transfomer decoder (Vaswani et al., 2017) in all models.

For the dataset construction method, we compare the method without zero anaphora resolution (**w/o ZAR**) and the one with zero anaphora resolution (**w/ ZAR**). For the decoder type, we compare the standard Transformer decoder and the one with a rewarding mechanism (**+ reward**).

For comparison, we use the response generation model that considers only the context as the Baseline. The Baseline model selects the topic entity using a heuristic method that regards the last entity in the context as the topic entity.

In decoding, we use sample-and-rank decoding (Adiwardana et al., 2020) for all models, including the Baseline. Each parameter is set at temperature T=1.0 and the number of response candidates N=50. For random sampling, top-k sampling and top-p sampling are applied, with k=40 and p=0.9. We also apply the bigram penalty (Paulus et al., 2018; Klein et al., 2017).

5.2 Evaluation Method

5.2.1 Topicality Estimation

We create the evaluation data for assessing topicality estimation using crowdsourcing.[7] First, we randomly select 57 dialogues from the test data of KUHCC and then extract 4,702 topic entity candidates along with the context using the method as in Section 3.2. Crowdworkers are shown the context and the topic entity candidates, and asked to select appropriate entities as the next topic from provided topic entity candidates (multiple choice is allowed).

---

[6]https://nlp.ist.i.kyoto-u.ac.jp/?ku_bert_japanese

[7]https://crowdsourcing.yahoo.co.jp/

	topicality estimation			response generation			
	P@1	R@1	R@3	PPL	BLEU-2/4	Natural	Topic
Baseline	0.561	0.440	0.777	24.40	7.14/0.88	2.46	2.52
Two-Stage w/o ZAR	0.625	0.491	0.803	**21.73**	7.22/**1.03**	2.33	2.52
Two-Stage w/o ZAR + reward	0.625	0.491	0.803	21.86	7.35/0.94	2.51	2.65
Two-Stage w/ ZAR	**0.658**	**0.523**	**0.838**	23.14	7.29/1.01	**2.75**	2.79
Two-Stage w/ ZAR + reward	**0.658**	**0.523**	**0.838**	23.14	**7.44**/1.01	2.71	**2.82**
End-to-End w/o ZAR	0.477	0.363	0.747	21.82	6.84/0.91	2.42	2.40
End-to-End w/o ZAR + reward	0.446	0.310	0.691	**21.73**	7.08/0.85	2.45	2.56
End-to-End w/ ZAR	0.580	0.449	0.776	21.86	6.87/0.91	2.56	2.59
End-to-End w/ ZAR + reward	0.605	0.479	0.801	21.87	7.11/0.82	2.45	2.63

Table 2: Results of evaluation of topicality estimation and response generation

All the entities selected by five or more workers are used as positive examples (= topic entities), and the rest are used as negative ones. Note that we remove the pairs of the context and the topic entity candidates for which no positive examples exist from the evaluation data. As a result, we obtained 741 positive examples and 3,219 negative examples as evaluation data.

We evaluate the models using the created evaluation data. We use P@1 and R@k as evaluation metrics. P@1 is the top-1 precision, and R@k is the top-k recall (k=1,3 in this paper).

5.2.2 Response Generation

We evaluate the models using both automatic metrics and human evaluations. For automatic metrics, we calculate perplexity (PPL) and BLEU-2/4 (Papineni et al., 2002) for test data of three corpora for fine-tuning. Perplexity measures the fluency of generated responses, and BLEU metrics measure the accuracy of generated responses in terms of lexical overlap with references.

In human evaluations, crowdworkers evaluate responses by their degree of agreement to the following questions referring to the method of Zhang et al. (2020), on a five-point Likert scale (1: completely disagree, 5: completely agree).

- **Naturalness (Natural)**: "Do you think the given response is natural as Japanese?"

- **Topic-Following (Topic)**: "Do you think the given responses follows the topic in the context?"

The crowdworkers are shown pairs of a context and a response. The input to the models for generating responses is 100 contexts randomly extracted from the test data of KUHCC, Each pair of context and response is rated by five crowdworkers.

5.3 Results and Analysis

5.3.1 Topicality Estimation

Table 2 shows the evaluation results of the topicality estimation. Two-Stage w/ ZAR achieves the best scores on both precision and recall.

For the dataset construction method, w/ ZAR is better than w/o ZAR for both Two-Stage and End-to-End models. These results suggest that restoring the word omissions help improve the accuracy of topicality estimation.

As for the decoder type, the Two-Stage model outperforms the End-to-End model on the whole. This may be because the Two-Stage model directly optimizes the topicality estimation, whereas the End-to-End model does both the topicality estimation and response generation.

5.3.2 Response Generation

The results of the automatic evaluation for the response generation are shown in Table 2. The Two-Stage and End-to-End models we proposed in this paper show lower perplexity than the Baseline. For BLEU metrics, Two-Stage outperforms Baseline for both BLEU-2/4, although End-to-End shows no improvement. This result suggests that while topicality estimation helps improve response generation, multi-task learning of topicality estimation and response generation does not improve response generation.

The results of the human evaluation for the response generation are also shown in Table 2. In terms of restoring omission of words, both Two-Stage w/ ZAR and End-to-End w/ ZAR achieve better Topic-Following score compared to the Baseline. This improvement indicates that restoring the omission of entities helps generate the topic-following responses. In addition, the Topic-

Following score further improves by adding a rewarding mechanism, which is the method to increase the generation probability of the topic entity.

As for the decoder type, the Two-Stage model is better than the End-to-End model in both Naturalness and Topic-Following. This difference may be due to the fact that the End-to-End did not learn well to reflect on the topic entity because only about 50% of all context-response pairs are labeled with some topic entities.

6 Conclusion

We proposed dialogue systems that explicitly consider topicality to generate responses following the topic. Both automatic and human evaluation results confirmed that the proposed Two-Stage model could generate more topic-following responses than the dialogue system that only considers context. In addition, by restoring the word omission in the response by zero anaphora resolution, topicality estimation was further improved, and it was also confirmed that the generated responses can better capture the topic.

On the other hand, the Naturalness score of the generated responses tends to be low overall in the human evaluation, and there is still room for improvement. We will work to improve the quality of the response generation part by using some pretraining models as future work.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977.

Kenneth Ward Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Givón. 1983. *Topic Continuity in Discourse: A quantitative cross-language study*, volume 3. John Benjamins Publishing Company.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3349–3358, Osaka, Japan. The COLING 2016 Organizing Committee.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A Deep Reinforced Model for Abstractive Summarization. In *In Proceedings of the 6th International Conference on Learning Representations*, Vancouver, BC, Canada.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural Responding Machine for Short-Text Conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2021. Empirical analysis of training strategies of transformer-based japanese chit-chat systems. *CoRR*, abs/2109.05217.

Yuto Takebayashi, Chu Chenhui, Yuki Arase†, and Masaaki Nagata. 2018. Word rewarding for adequate neural machine translation. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 14–22, Brussels. International Conference on Spoken Language Translation.

Arseny Tolmachev, Daisuke Kawahara, and Sadao Kuro-hashi. 2018. Juman++: A Morphological Analysis Toolkit for Scriptio Continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 54–59, Brussels, Belgium. Association for Computational Linguistics.

Nobuhiro Ueda, Daisuke Kawahara, and Sadao Kuro-hashi. 2020. BERT-based Cohesion Analysis of Japanese Texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1323–1333, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Oriol Vinyals and Quoc V. Le. 2015. A Neural Conversational Model. In *Proceedings of the International Conference on Machine Learning, Deep Learning Workshop*.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic Aware Neural Response Generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3351–3357. AAAI Press.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Shuying Zhang, Tianyu Zhao, and Tatsuya Kawahara. 2020. Topic-relevant Response Generation using Optimal Transport for an Open-domain Dialog System. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4067–4077, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wayne Zhao, Jing Jiang, Js Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing Twitter and Traditional Media Using Topic Models. volume 6611/2011, pages 338–349.