More Informative Dialogue Generation via Multiple Knowledge Selection

Anonymous ACL submission

Abstract

Knowledge-grounded dialogue generation is a task of generating a fluent and informative response based on both dialogue context and a collection of external knowledge. There is a lot of noise in the knowledge pool, and appropriate knowledge selection plays an important role. Existing methods can only select one piece of knowledge to participate in the generation of the response, which inevitably loses some useful clues contained in the discarded candidates. In this work, we propose MSEL, a novel knowledge selector which could select multiple useful knowledge. MSEL takes the dialog context and knowledge pool as inputs and predicts a subset of knowledge pool in sequence-to-sequence manner. MSEL is easy to implement and can benefits from the generative pre-trained language models. Empirical results on the Wizard-of-Wikipedia dataset indicate that our model can significantly outperforms state-of-the-art approaches in both automatic and human evaluation.

1 Introduction

Open domain human-machine conversation system have drawn increasing attention from the research community of artificial intelligence and natural language processing due to the rapid development of sequence to sequence modeling technology (Sutskever et al., 2014; Vaswani et al., 2017). However, if we directly apply some commonly used sequence generation models to dialog systems, the generated response will be very simple and generic, and its informativeness is far from human performance (Li et al., 2016; Lian et al., 2019). To bridge the gap, researchers begin to study how to ground open domain dialogues by external knowledge, which could be obtained from some knowledge bases such as Wikipedia by information retrieval technology.

Knowledge-grounded dialogue is a task of generating an informative response based on both dialogue context and a group of external knowledge. Dinan et al. (2019) proposed to decompose this task into two subtasks: knowledge selection and response generation, and proposed a benchmark dataset named Wizard-of-Wikepedia (WoW).

Context	What do you know about Ireland?
Knowl. Pool	1. Ireland (Ulster-Scots) is an island in the North Atlantic. 2. Northern Ireland (Ulster-Scots) is a ... 3. William Wentworth-Fitzwilliam, 4th ... 4. The Troubles was an ethno-nationalist ... 5. The population of Ireland was about 6.6 million, ranking it the second-most populous island in Europe. ... ...
Refenence Response	Are you referring to the island in the North Atlantic? It has a population of 6.6 million people. Why do you ask?

Table 1: An example from WoW. There are 40 pieces of knowledge and we only show a part.

We manually analyzed the WoW dataset and found that in many cases, just selecting a piece of knowledge is not enough. For example, as shown in Table 1, in order to generate the gold response, model need to catch knowledge #1 and #5 simultaneously. However, most existing methods (Lian et al., 2019; Dinan et al., 2019; Kim et al., 2020; Zhao et al., 2020; Zhan et al., 2021; Meng et al., 2021) can only select one knowledge with the highest confidence from the candidate knowledge to participate in dialogue generation, while abandoning other knowledge with low confidence but possibly containing useful information.

To overcome the challenge, we propose using a novel and simple sequence-to-sequence (Seq2Seq) framework to generate the a set of useful knowledge directly. On the source side, the model inputs the concatenation of dialogue history and knowledge pool, and on the target side, the model generates the knowledge pointer index sequence. By converting the knowledge selection task into a Seq2Seq

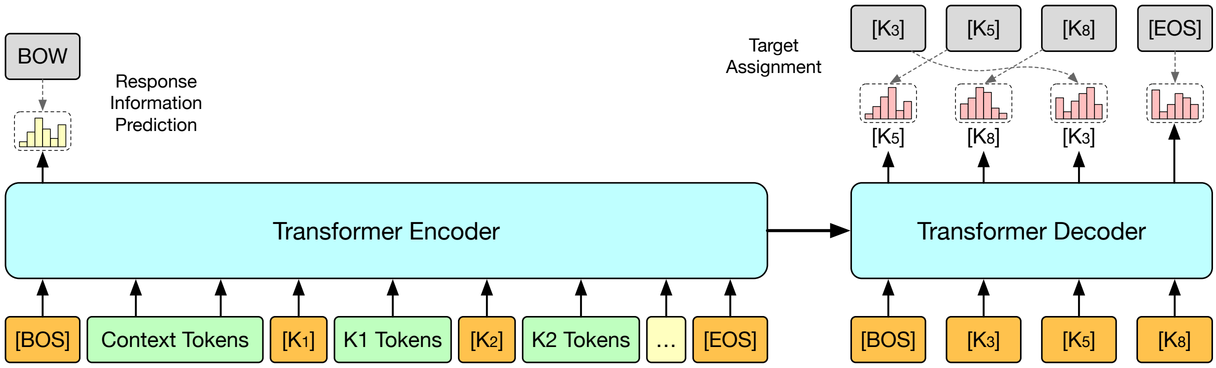

Figure 1: The architecture of our MSEL.

generation task, we can smoothly use the generative pre-training language model such as BART (Lewis et al., 2020) to enhance our model.

Our contribution can be summarized as follows: 1) We observe that in knowledge-grounded dialogue, in many cases, multiple pieces of external knowledge need to be selected; 2) We formulate the knowledge selection task in knowledge-grounded dialogue as a sequence prediction problem, and propose MSEL, a simple but effective approach; 3) Our proposed method yields new state-of-the-art performance on Wizard-of-Wikipedia (Dinan et al., 2019) dataset.

The source code of MSEL will be available for public.

2 Approach

Given dialogue context $U = \{u_1, ..., u_m\}$ and the external knowledge pool $\mathcal{K} = \{K_i\}_{i=1}^{s}$, the goal of our model is to choose a subset of $\mathcal{K}$ most relevant to the generation of response $R = \{r_1, ..., r_n\}$. The architecture of our model is shown in Figure 1.

2.1 Target Construction

We focus on the scenarios where ground-truth knowledge is unknown, so we does not use the knowledge label provided by benchmark datasets during training. Intuitively, responses from humans carry clues to relevance of the knowledge candidates (Zhao et al., 2020), so the knowledge sentence that promotes the flow of conversation usually has a higher textual similarity with the gold response. Based on this assumption, we construct a set of pseudo-label (denoted as $\mathcal{D}$) by

$$\mathcal{D} = \{K_{a_1}, K_{a_2}, ..., K_{a_z}\} \tag{1}$$

$$= \underset{\hat{\mathcal{D}} \subseteq \mathcal{K}}{\arg\max}\, Sim(\phi(\hat{\mathcal{D}}), R) \tag{2}$$

where $\phi(\cdot)$ refers to the concatenation operator and Sim refers to unigram F1-score. The model is required to predict the knowledge index in $\mathcal{D}$.

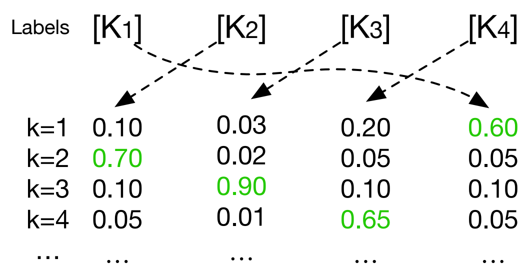

Figure 2: An example of set prediction paradigm.

2.2 Seq2Seq for Knowledge Selection

As shown in Figure 1, we formulate the knowledge set selection task in a generative way. For a sample, we concatenate the dialog context with all knowledge in $\mathcal{K}$ as source input:

$$X = \{[\mathrm{BOS}], U, ..., [\mathrm{K}_i], K_i, ..., [\mathrm{EOS}]\} \tag{3}$$

where [BOS] and [EOS] are two control tokens which indicates begin of sentence and end of sentence, and $[\mathrm{K}_i]$ $(1 \leq i \leq s)$ is a special token used to mark the position of knowledge in the input sequence. X is pass to a Transformer encoder to get the context-aware representation:

$$\mathbf{H} = \mathrm{TFEncoder}(X) \tag{4}$$

Decoder is to get the probability distribution for each step $p_t = p(y_t|X, Y_{<t})$, where Y refers special token sequence $\{[\mathrm{K}_{a_1}], ..., [\mathrm{K}_{a_z}], [\mathrm{EOS}]\}$.

As shown in Figure 2, sometimes the decoder does not predict the knowledge indexes in the order of the ground-truth sequence, so we need to reassign Y. Inspired by the assigning problem in operation research (Kuhn, 2010) and motivated by (Sui et al., 2020), we propose a set prediction loss that can produce an optimal bipartite matching between predicted and ground-truth labels. Specifically, we need to optimize the following problems

$$\pi^* = \arg\max_{\pi \in \Pi(z)} \sum_{t=1}^{z} p(y_{\pi(t)}|X, Y_{<t}) \quad (5)$$

$$Y^* = \{y_{\pi^*(1)}, y_{\pi^*(2)}, ..., y_{\pi^*(z)}, y_{z+1}\} \quad (6)$$

where $\Pi(z)$ is the space of all z-length permutations. The process of computing bipartite matching loss is divided into two steps: finding an optimal matching and computing the loss. This optimal assignment π^* is computed in polynomial time via the Hungarian algorithm. We define the loss as the $\mathcal{L}_{\text{NLL}}$.

Since knowledge labels are annotated by responses, it is hard to select knowledge only depend on discourse context, which is a phenomenon similar to the posterior collapse problem (Zhao et al., 2017a). In order to enhance the prior selection with the necessary posterior information, we add a submodule that predicts the response information (Chen et al., 2020; Li et al., 2020). Specifically, we feed the hidden states of [BOS] to a small network to generate the response information in bag-of-words (BOW) format (Zhao et al., 2017b):

$$\hat{\mathbf{I}} = \sigma\left(\text{MLP}(\mathbf{h}_{[\text{BOS}]})\right) \in \mathbb{R}^{|V|} \quad (7)$$

We supervise this module by an addition loss

$$\mathcal{L}_{\text{BOW}} = -\frac{1}{|V|} \sum_{w \in \mathbf{I}} \log(\hat{\mathbf{I}}_w) \quad (8)$$

where $\mathbf{I}$ is the BOW vector of ground-truth response. The total loss of the knowledge selector can be expressed as: $\mathcal{L} = \mathcal{L}_{\text{NLL}} + \mathcal{L}_{\text{BOW}}$.

2.3 Response Generation

Although there are various methods studying how to improve the generation quality based on the selected knowledge, here, we simply follow the baselines (Kim et al., 2020; Zhan et al., 2021) that concatenate $\mathcal{K}_{sub}$ with the dialog context and feed to a Transformer to generate response token by token.

3 Experiments

3.1 Dataset and Metrics

Dataset We use Wizard-of-Wikipedia (WoW) (Dinan et al., 2019) for our experiments. WoW contains over 20k social chat conversations between two agents, and the two participants are not quite symmetric: one will play the role of a knowledgeable expert (i.e., wizard) while the other is a curious learner (i.e., apprentice). Each wizard turn is associated with ∼40 pieces retrieved from the Wikipedia and each piece contains ∼30 words, and most of them are noise. The test set is split into two subsets, test seen and test unseen. The difference between the two is that the former contains some topics that overlap with the training set.

Metrics We use both quantitative evaluation and human judgements in our experiments. Specifically, we choose corpus-level BLEU (Papineni et al., 2002), Dist (Li et al., 2016) and embedding-based indicators [1] (average, extrema and greedy) as automatic evaluation metrics. Response with higher embedding-based score and BLEU is closer to the ground-truth and more fluent. Dist is a reference-free metric to evaluate the diversity, and response with higher Dist carry more information. Besides quantitative evaluation, we also recruit three human annotators to do qualitative analysis on response quality. We randomly sample 100 samples, and each sample contains the conversation history, response, and external knowledge set. The annotators then judge the quality of the responses from three aspects, including context coherence, language fluency and response informativeness, and assign a score in {0, 1, 2} to each response for each aspect. Each response receives 3 scores per aspect, and the agreement among the annotators is measured via Fleiss' kappa (Fleiss, 1971).

3.2 Baselines

We compare our MSEL with several competitive baselines, including (1) **TMN**: The end-to-end Transformer with memory mechanism (Dinan et al., 2019); (2) **PostKS**: A LSTM-based model that leverages posterior knowledge distribution to enhance the selector (Lian et al., 2019); (3) **SKT**: A sequential latent knowledge selection model which keeps track the selection history over conversation flow (Kim et al., 2020); (4) **SKT-KG**: A transition model that tracking the knowledge selection history with CRF (Zhan et al., 2021); (5) **Blender-KG**: Blender is a Transformer model which pre-trained on massive conversations (Roller et al., 2021). Similar to Cui et al. (2021), we concatenate dialog context and knowledge pool as the input of Blender; (6) **KnowledGPT**: A knowledge-grounded dialogue

[1] https://github.com/Maluuba/nlg-eval

Model	Test Seen					Test Unseen				
	BLEU-1/2	Dist-1/2	Avg.	Ext.	Gre.	BLEU-1/2	Dist-1/2	Avg.	Ext.	Gre.
TMN	16.7/6.7	3.6/11.2	48.2	40.3	44.2	14.7/4.8	2.1/16.2	42.3	35.7	38.5
PostKS	17.2/7.0	5.7/21.8	53.3	39.2	45.1	15.6/5.4	2.9/15.2	44.2	38.7	40.4
SKT	18.9/7.6	7.3/26.5	54.0	43.6	51.2	15.9/6.1	2.3/16.6	42.0	39.2	43.7
SKT-KG	20.6/7.4	7.8/28.3	59.7	**48.8**	54.3	16.3/**7.0**	3.6/16.8	52.8	41.0	45.4
Ours	**21.4/9.5**	**7.4/37.0**	**86.2**	45.0	**66.5**	**19.2/7.0**	**5.1/27.4**	**85.9**	**42.1**	**65.3**
Blender-KG	23.6/11.8	7.9/34.3	86.9	45.8	68.0	22.3/10.0	5.3/24.1	86.8	44.5	67.1
KnowledGPT	25.7/14.2	8.9/36.2	86.1	**46.2**	68.2	24.5/**12.8**	6.0/23.8	87.0	45.3	67.4
Ours + BART + GPT-2	**26.0/14.4**	**9.5/38.7**	**87.0**	46.1	**68.4**	**24.6/12.8**	**6.3/26.5**	**87.1**	**45.7**	**67.6**

Table 2: Automatic evaluation results on Wizard-of-Wikipedia.

model based on reinforcement learning (Zhao et al., 2020). SKT and SKT-KG perform BERT (Devlin et al., 2019) as selector, and use Transformer decoder as the response generator. KnowledGPT use BERT to select knowledge and use GPT-2 to generate response. Some other models do not release the entire source code and use different metrics with us, so we do not make comparisons with them.

3.3 Implementation Details

The number of Transformer layers is set to 12. We limits the maximum length of input to 1024 tokens, and truncated the redundant sequence. To retain as much useful information as possible, we trained a simple ranking model to rank the knowledge in the pool, so that the noise knowledge is arranged at the end of the sequence as much as possible.

3.4 Results

Table 2 reports the results on automatic evaluation, and the dividing line of the table is used to distinguish whether the response generator of the model uses a pre-trained language model. We can observed from the upper part of the table that our model surpasses all baselines in most metrics without leveraging pre-trained language model. Our model has a great advantage in Dist-n, which means that our multiple knowledge selection mechanism can incorporate more information into the response. In order to test the performance of MSEL that equipped with pre-trained language models, we use BART to initialize the parameters of the selector, and use GPT-2 as the response generator. We can conclude from the bottom half of Table 2 that our model outperforms Blender-KG and KnowledGPT.

The results of human evaluation are shown in Table 3, which also indicates that our model can generate more informative response than baselines.

Model	Coherence	Fluency	Info.	Kappa
Transformer	1.63	1.70	1.34	0.59
SKT	1.72	1.75	1.63	0.58
MSEL	**1.73**	**1.76**	**1.71**	0.59

Table 3: Human evaluation results on WoW test seen. Transformer refers to no knowledge access.

3.5 Ablation Study

We conduct ablation experiments on WoW, and the results are shown in Table 4. To verify whether the improvements are owing to the multiple selection mechanism, we try not use knowledge, use all knowledge and use only one knowledge to train the generator. It can be seen from the results that this mechanism is effective. Although use all knowledge seems more informative, noise in knowledge pool can mislead the topic of the dialog, which results a lower BLEU-1/2. We also tried to remove some of the components, such as $\mathcal{L}_{\text{BOW}}$ and the target assignment, and the performance decreased.

Model	Test Seen		Test Unseen	
	B-1/2	D-2	B-1/2	D-2
MSEL	**21.4/9.5**	**37.0**	**19.2/7.0**	27.4
w/o Knowledge	17.3/6.6	26.7	16.9/5.5	18.7
Use all Knowl.	20.7/8.8	34.6	17.7/5.5	**28.5**
Sel. One Kno.	20.7/8.3	31.4	18.3/6.4	25.3
w/o $\mathcal{L}_{\text{BOW}}$	21.2/9.4	36.3	19.2/6.8	26.9
w/o Target Assi.	21.3/9.3	36.5	19.0/6.9	26.7

Table 4: Ablation Study results on WoW.

4 Conclusion

In this paper, we explore knowledge-grounded dialogue generation with multiple knowledge selection, and propose a sequence-to-sequence framework to tackle this task. Evaluation results on WoW indicate that our model can generate more informative response, and achieves satisfied performance.

References

Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020. Bridging the Gap between Prior and Posterior Knowledge Selection for Knowledge-Grounded Dialogue Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3426–3437.

Leyang Cui, Yu Wu, Shujie Liu, and Yue Zhang. 2021. Knowledge Enhanced Fine-Tuning for Better Handling Unseen Entities in Dialogue Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2328–2337. Association for Computational Linguistics.

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)*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *International Conference on Learning Representations*.

Joseph Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential Latent Knowledge Selection for Knowledge-Grounded Dialogue. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Harold W. Kuhn. 2010. The hungarian method for the assignment problem. In Michael Jünger, Thomas M. Liebling, Denis Naddef, George L. Nemhauser, William R. Pulleyblank, Gerhard Reinelt, Giovanni Rinaldi, and Laurence A. Wolsey, editors, *50 Years of Integer Programming 1958-2008 - From the Early Years to the State-of-the-Art*, pages 29–47. Springer.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *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.

Linxiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-Resource Knowledge-Grounded Dialogue Generation. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to Select Knowledge for Response Generation in Dialog Systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5081–5087.

Chuan Meng, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tengxiao Xi, and Maarten de Rijke. 2021. Initiative-Aware Self-Supervised Learning for Knowledge-Grounded Conversations. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 522–532. ACM.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

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, EACL 2021, Online, April 19 - 23, 2021*, pages 300–325. Association for Computational Linguistics.

Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, Xiangrong Zeng, and Shengping Liu. 2020. Joint entity and relation extraction with set prediction networks. *CoRR*, abs/2011.01675.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural*

Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pages 5998–6008.

Haolan Zhan, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Yongjun Bao, and Yanyan Lan. 2021. Augmenting Knowledge-grounded Conversations with Sequential Knowledge Transition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5621–5630, Online. Association for Computational Linguistics.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017a. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017b. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-Grounded Dialogue Generation with Pre-trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390. Association for Computational Linguistics.

A Appendix

A.1 More Implementation Details

We implement our model with PyTorch framework. The initial word embedding size and hidden size is set to 768, and the vocabulary size is set to 50265. The hidden size in FFN of Transformer is set to 3072. The batch size is set to 32. We run all models on the Quadro RTX 8000 GPU.

A.2 Case Study

Table 5 shows a case from Wow test seen, from which we can see that MSEL selects two knowledge and generates more informative response than Transformer and SKT. Transformer (without knowledge accessing) can generate a relatively informational response because it has seen similar samples during training.

Context	New York City has always fascinated me, have you ever been there?
Reference	I haven't, but I don't know if I do want to go or not lol. It's the most populated city in America, so I don't know if I'm up to so many people!
MSEL Selected Knowledge	1. The City of New York, often called ... ... most populous city in the United ... 2. A global power city, ... has been described as the cultural, financial, and media capital of the world ... ...

Transformer: Well, in 2016 there were five boroughs, Brooklyn, Queens, Queens, Manhattan, and The Bronx.
SKT: Yes, it is the most populous city in the America.
MSEL: I have! I grew up in New York City, the most densely populated major city in the US, and has been described the cultural, financial, and media capital of the world.

Table 5: A case from test seen of WoW.