For all Reviewers: Thank you for the valuable comments that help us improve the work.

For Reviewer #1: 1. *Effect of UniLM:* We observe obvious performance drop when using fine-tuned UniLM with fixed top-1 retrieved knowledge (-parameterized posterior in Table 3). Following your suggestion, we add an experiment in which we train our model with all parameters randomly initialized, and the results are: $42.9, 18.6, 44.5, 18.5, 44.9, 15.7, 41.3, 16.3, 51.1, 11.8$ (aligned with the columns of Table 1). Therefore, the performance of ZRKGC with UniLM initialization slightly improves but the major improvement comes from our method.

2. *Knowledge selection in the test time:* In case there is any misunderstanding, let us clarify our motivation again. We propose a method to automatically connect Reddit and Wikipedia in replacement of expensive human-annotated dataset (e.g. Wizard) to train a knowledge-grounded generation model. Before us, models are trained on the human-annotated datasets, and are tested with the hold-out data annotated in the same way. Therefore, to draw a fair comparison, we keep the same evaluation procedure with the existing models. Another issue that needs clarification is that in test time the knowledge selection model $p(y|C, Z)$ is responsible for selecting K knowledge sentences from all M knowledge sentences(M>=K) to prevent the length of input from exceeding the maximum length of UniLM while previous work [18][25] focuses on selecting top-1 knowledge. Our model performs implicit knowledge selection on the input K knowledge sentences (concatenated in a sequence) in an end-to-end way like DRD [52]. (i): Yes, for ITDD we do not explicitly select knowledge and simply concatenate all the knowledge candidates into a sequence. (ii): All the models use the original WoW setting with an average of 60~70 knowledge sentences in each dialogue turn. In both human evaluation and case study, all models generate responses with all knowledge sentences. We do not list all knowledge sentences in Table 4 and Table 5, because that will make them hard to read; (iii): For our model, the proportion of the GT-knowledge in the input K knowledge sentences on WoW seen and WoW unseen are 37.7% and 37.4% respectively.

3. *Supervision of knowledge selection:* (i): We have tried varying the number of retrieved knowledge in [1,5,10,15,20,25,30] before. F1 on the validation set increases until the number of knowledge reaches 10, but stays stable when the number increases from 10 to 30. Finally, to speed up training, we use the number 10. We train our model by randomly choose Z_k from the top 10 retrieved knowledge from the Lucene retriever, and the results are: $43.0, 16.7, 44.8, 16.5, 44.6, 12.2, 42.1, 12.9, 53.0, 9.9$ (aligned with the columns of Table 3). Therefore, both randomly sampling and keeping only top-1 knowledge (-parameterized posterior in Table 3) will cause dramatic performance drop, indicating the effectiveness of our method; (ii): The results of "without knowledge loss" are: $44.2, 18.3, 45.9, 17.9, 45.5, 14.6, 43.5, 14.9, 53.8, 12.0$ (aligned with the columns of Table 3). In test time, knowledge selection module is mainly to shorten the input sequence of knowledge candidates, so the performance drop is not significant. (iii): Your understanding is correct. But in practice the Lucene retriever selects knowledge based on responses while the model selects knowledge with only access to contexts and knowledge candidates, so it is very hard for the model to select the same knowledge sentence with Lucene retriever.

4. *Choice of baselines:* We choose DRD instead of SKT [18] and PostKS [25] because (i) human annotations on knowledge selection are crucial to the performance of SKT and PostKS and such annotations (in sentence level) are not available in Topical-Chat and CMU_DoG; and (ii) both SKT and PostKS perform worse than DRD on Wizard [18][25][52]. In spite of this, for your reference, we implement SKT with heuristics on Topical-Chat and CMU_DoG (pseudo supervision created by selecting GT-knowledge using Sim(.,.) with the response), and the results are: $52.0, 19.3, 81.4, 16.1, 25.1, 17.0, 35.6, 14.8, 41.9, 9.6$ (aligned with the columns of Table 1), which are worse than our method. Besides optimization using generalized EM, our model introduces another variable Z_α to dynamically adapt to candidates in different quality while SKT and PostKS assume there always exists GT-knowledge in their candidates. For REALM, the notification date of ICML 2020 is quite close to the submission date of NeurlPS 2020. A thorough discussion about these three works will be presented in the final version.

For Reviewer #2: We will follow your suggestions on the improvement of clarity in the final version. We have tried increasing the search space of knowledge with details described in **For Reviewer #1:**3(i). Increasing the search space will cost more GPU memory and training time, and thus it is not easy to scale.

For Reviewer #3: During test time, the knowledge candidates are already prepared in test dataset, we only need to tailor the knowledge candidates to meet the capacity constraint of UniLM by using knowledge selector $p(y = 1|C, Z)$.

For Reviewer #4: We will follow your suggestions on the improvement of clarity in the final version. "Zero-Resource" in our paper specifically refers to training model without crowd-sourced knowledge-grounded dialogues (e.g. Wizard, Topical Chat) rather than completely without any knowledge resources (e.g. Wikipedia). Existing knowledge-grounded dialogue generation models are all trained on crowd-sourced knowledge-grounded dialogue datasets. Such datasets are enormously expensive to collect, and thus are small in size (Wizard contains only about 150k utterances, for example). We explore a way to train such a model from easy-to-collect datasets (e.g. Reddit and Wikipedia) in the paper, so our training data could have up to about 2600k utterances. The generalization ability of our model originates from the ability of our proposed method to break through the limitation of training data size.