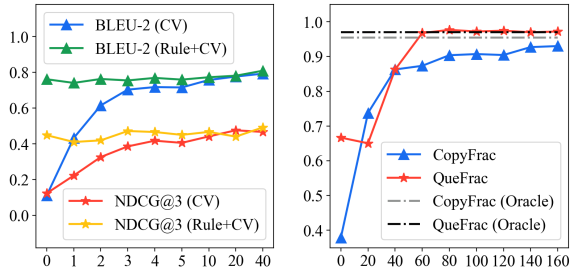

(a) Training Sessions (b) Training Steps

Figure 2: Performances of GPT-2 with different fine-tuning amounts: conversational sessions with manual rewrites (a) and fine-tuning steps (b). The Y-axes show the corresponding metric in (a) and (b).

In the few-shot setting, GPT-2 trained with CV already outperforms the best CAsT auto runs, pgbert and CFDA. Together with weak supervision data, Rule-Based + CV or Self-Learn + CV improves the state-of-the-art by 10+%. The improvement is mainly attributed to better query rewriting: our simple BERT (base) ranker, when using Oracle queries, is less effective than pgbert and CFDA teams' manual runs; they obtained 0.57+ NDCG@3, compared to ours 0.544 [1]. The BLEU scores correlate well with NDCG—better query rewriting leads to better search accuracy. Our query rewriter also maintains a stable accuracy in later turns, as shown in Fig. 1b, which indicates that our rewriter effectively captures the multi-turn context as the conversation proceeds.

Surprisingly, GPT-2 (CV) provides effective rewrites when only cross validated on 50 CAsT sessions; Rule-Based, in the zero-shot setting, is on par with best TREC CAsT automatic runs (Fig. 1a shows their individual effectiveness). In comparison, directly applying (GPT-2 Raw) or only fine-tuning using ad hoc sessions (MARCO Raw) yield sub-par results. It is impressive that the pre-trained transformer can learn conversational query rewriting, a challenging task for previous techniques, in such a data efficient manner.

5.2 Few-Shot Study

This study further investigates GPT-2's capability of generalization.

How Few Shot? Fig. 2a shows GPT-2 fine-tuned with fewer sessions, with or without weak supervision. Exceptionally, GPT-2 learns to generate reasonable query rewrites with *only three conversational sessions or 30 manual labels*; it matches best CAsT auto runs with as few as 10 sessions.

What is Learned? It is unlikely that GPT-2 learns the discourse phenomena from just three sessions. They are likely to be captured in pre-training since the non-pre-trained GPT-2 does not outperform substantially random guess, as in Table 2.

We hypothesize that GPT-2 only needs to learn the "syntax" of the rewriting task during fine-tuning: to generate questions and to replace pronouns with or add concepts mentioned in previous turns. Fig. 2b plots the fraction of questions (QueFrac) in GPT-2 (CV) rewrites, indicated by question words, and the percentage of new words being copied from previous queries (CopyFrac), at different fine-tuning steps. GPT-2 adapts to query rewriting very quickly with very little fine-tuning. Our effectiveness perhaps is more from properly "unleashing" the language understanding power already in the pretrained language model.

Table 3: GPT-2 Query Rewrites on CAsT Topic 31 and 64.

Q_6	What causes **throat cancer**?
Q_7	What is the first sign of it?
Q_8	Is it the same as **esophageal cancer**?
Q_9	What's the difference in their symptoms?
Oracle	What's the difference in **throat cancer and esophageal cancer's** symptoms?
Output	What's the difference between **throat cancer and esophageal cancer**?
Q_1	What are the types of **pork ribs**?
Q_2	What are baby backs?
Q_3	What are the differences with spareribs?
Q_4	What are ways to **cook** them?
Q_5	How about on the bbq?
Oracle	How **do you cook pork ribs** on the bbq?
Output	How about on the bbq?

5.3 Case Study

Table 3 provides two examples from GPT-2 (Rule-based + CV). We found it surprising that in the first case, GPT-2 accurately resolves the group coreference from "their" to two cancer types, with one of the two from three turns ago. The second example presents a common error made by our rewriter: it fails to add proper context because it is not clear what the context the term "about" refers to.

6 CONCLUSION

This work demonstrates the effectiveness of GPT-2 for conversational query rewriting. Fine-tuned using weak supervision data generated by rules or a handful of manual rewriting labels, our GPT-2 query rewriter is able to create new state-of-the-art on the TREC CAsT conversational search benchmark—outperforming previous methods including query expansion, contextual ranking, and coreference resolution, many of which use large-scale pre-trained models and deep neural networks.

ACKNOWLEDGEMENTS

This work is supported by the National Key Research and Development Program of China (No. 2018YFB1004503) and the National Natural Science Foundation of China (NSFC No. 61732008, 61532010).

REFERENCES

[1] Jeff Dalton, Chenyan Xiong, and Jamie Callan. 2019. CAsT 2019: The Conversational Assistance Track Overview. In *TREC 2019*. NIST.
[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL 2019*.
[3] R. Nogueira and K. Cho. 2019. Passage Re-ranking with BERT. *ArXiv* abs/1901.04085 (2019).
[4] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).
[5] Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2020. Question Rewriting for Conversational Question Answering. *ArXiv* abs/2004.14652 (2020).