[47] Jian Wang and Wenjie Li. 2021. Template-guided Clarifying Question Generation for Web Search Clarification. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3468–3472.

[48] Sean Welleck, Kianté Brantley, Hal Daumé Iii, and Kyunghyun Cho. 2019. Non-monotonic sequential text generation. In *International Conference on Machine Learning*. PMLR, 6716–6726.

[49] Julia White, Gabriel Poesia, Robert Hawkins, Dorsa Sadigh, and Noah Goodman. 2021. Open-domain clarification question generation without question examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 563–570. https://doi.org/10.18653/v1/2021.emnlp-main.44

[50] Jingjing Xu, Yuechen Wang, Duyu Tang, Nan Duan, Pengcheng Yang, Qi Zeng, Ming Zhou, and Xu Sun. 2019. Asking Clarification Questions in Knowledge-Based Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 1618–1629. https://doi.org/10.18653/v1/D19-1172

[51] Liu Yang, Hamed Zamani, Yongfeng Zhang, Jiafeng Guo, and W Bruce Croft. 2017. Neural matching models for question retrieval and next question prediction in conversation. *arXiv preprint arXiv:1707.05409* (2017).

[52] Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-shot generative conversational query rewriting. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 1933–1936.

[53] Hamed Zamani and Nick Craswell. 2020. Macaw: An extensible conversational information seeking platform. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2193–2196.

[54] Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*. 418–428.

[55] Hamed Zamani, Johanne R Trippas, Jeff Dalton, and Filip Radlinski. 2022. Conversational information seeking. *arXiv preprint arXiv:2201.08808* (2022).

[56] Maosen Zhang, Nan Jiang, Lei Li, and Yexiang Xue. 2020. Language Generation via Combinatorial Constraint Satisfaction: A Tree Search Enhanced Monte-Carlo Approach. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 1286–1298. https://doi.org/10.18653/v1/2020.findings-emnlp.115

[57] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th acm international conference on information and knowledge management*. 177–186.

[58] Ziliang Zhao, Zhicheng Dou, Jiaxin Mao, and Ji-Rong Wen. 2022. Generating Clarifying Questions with Web Search Results. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 234–244.

A Human Annotation Guideline

In this task, imagine you are the user who unintentionally asks our search system an ambiguous search query (imagine they are using Google or talking to Siri) in a conversation. To better understand the intention of your query, our system asks a clarifying question to you. And your task is to judge if this clarification question is natural and useful.

For example, the user asks "Tell me about defender". The query is ambiguous because the word "defender" can refer to a personality type coded as ISFJ, a TV series "The Defender", a vehicle named "Defender", or a video game named "Defender". In order to know whether the user is asking about the TV series, the search system asks a clarifying question "Are you interested in a television series?"

Another example can be the user asks "Tell me information about computer programming." Different from the last example, this query is NOT ambiguous because of the term "computer programming" is ambiguous, but because "computer programming" is a general concept, and there can be multiple search directions. For example, the user can be looking for computer programming jobs, computer programming languages, computer programming courses, or the history of computer programming. To confirm whether the user is looking for computer programming courses, the system asks a clarifying question "Are you looking for a course in computer programming?"

In general, ambiguous queries have many possible "facets". For example, "TV series" is one possible facet in the "defender" example, and "course" is one possible facet in the "computer programming" example. Our system generates these questions based on the ambiguous query and one possible facet.

Your goal is to evaluate the clarifying question asked by our system, in terms of its Naturalness and Usefulness (Please read explanation below). Besides the query, facet, and generated question, you will also get a human-written question as your reference. You can assume the human-written question is always good in both naturalness and usefulness.

Explanation of Naturalness:

The Naturalness of a question is whether the question is fluent, grammatical, and easy to understand. Your goal is to give each question "Good", "Fair", or "Bad" in terms of its naturalness. Good naturalness means the question is fluent and like our daily language. Fair naturalness means although not grammatically perfect or contains noise, the question can still be understood with efforts. Bad naturalness means the question is incomplete, hard to understand or the generated sentence is not a question.

Here are some examples with explanations:

Example 1

Query: "Tell me about defender"

Facet: "television series"

Reference: "are you interested in the television series defender"

Good naturalness questions:

"are you interested in a television series" (Almost the same as reference)

"do you need to be in the team" (fluent and easy to understand, although not meaningful)

Fair naturalness questions:

"do you want to know television series" (A little weird but understandable)

Bad naturalness questions:

"television series, etc." (Not a question, and the sentence is incomplete)

"would you like to know more about" (the question is incomplete)

Example 2

Query: "Tell me information about computer programming."

Facet: "courses"

Reference: "are you interested in coding courses online"

Good naturalness questions:

"are you looking for a course in computer programming" (Almost the same as reference)

"do you need to have courses in computer science" (Almost the same as reference)

"would you like to tell me about it" (fluent and easy to understand, although not meaningful)

Fair naturalness questions:

"do you want to coursework in computer programming" (sound strange/ungrammatical, but understandable on a second thought)

"do you want to know what is going on with your courses" (fluent but unlike daily language)

Bad naturalness questions:

"do you need to know" (Not a complete sentence)

Explanation of Usefulness:

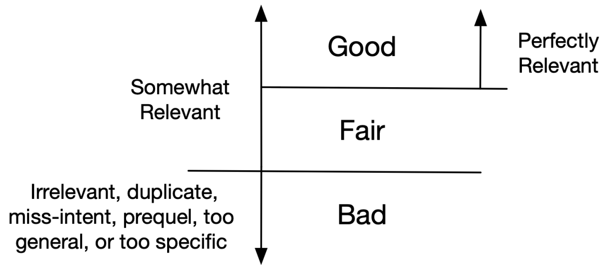
Good
Perfectly Relevant
Somewhat Relevant
Fair
Irrelevant, duplicate, miss-intent, prequel, too general, or too specific
Bad

Figure 2: Decision Boundaries for Usefulness

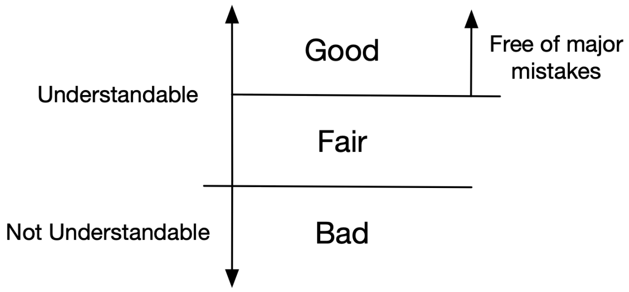

Figure 1: Decision Boundaries for Naturalness

A useful question should:

(1) Clarify the query if answered by the user.
(2) Not be a duplicative question of the query or ask for prequel information.
(3) Not miss the intent of the query.
(4) Not be too general or over-specific.

Your goal is to give each question "Good", "Fair", or "Bad" in terms of its usefulness. A Good usefulness questions is a perfect reflection of the facet and make the query easier to answer. A Fair usefulness questions is weakly relevant to the query and facet, but not completely irrelevant. A Bad usefulness question can be completely irrelevant, duplicative, miss-intent, prequel, too general or over-specific.

A question can be natural but not useful. Please see examples below.

Example 1

Query: "Tell me about defender"

Facet: "television series"

Reference: "are you interested in the television series defender"

Good usefulness questions:

"are you interested in a television series" (Almost the same as reference)

"do you want to know television series" (not perfectly natural but useful)

Fair usefulness questions:

"would you like to see a television series based on your work" (although "based on your work" is weird, user could still answer the question as "yes I am referring to the TV series defender")

Bad usefulness questions:

"do you need to be in the team" (not meaningful for the query)

"television series, etc." (Not a question, and the sentence is incomplete, user cannot answer it)

Example 2

Query: "Tell me information about computer programming."

Facet: "courses"

Reference: "are you interested in coding courses online"

Good usefulness questions:

"are you looking for a course in computer programming" (Almost the same as reference)

"do you need to have courses in computer science" (not natural but useful)

"do you want to coursework in computer programming" (not natural but useful)

Fair usefulness questions:

"do you want to know what is going on with your courses" (weakly relevant to the facet)

Bad usefulness questions:

"do you need to know" (Not a complete sentence)

"would you like to tell me about it" (completely irrelevant)