Table 2: Model performances evaluated on full question and reference. ZSFC models are our zero-shot facet-constrained method with different question rankers. Bolded numbers indicate the best-performing model of the column.[1]

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	Coverage
Q-GPT-0	4.31	2.49	1.79	1.45	5.21	5.60	1.33
QF-GPT-0	5.49	3.21	2.30	1.84	7.50	8.32	7.53
Prompt-0	5.25	3.00	2.13	1.70	6.42	7.38	7.89
Template-0	21.27	15.26	11.99	9.96	27.58	24.04	6.86
Subject-constrained	29.16	20.99	16.52	13.68	28.59	34.78	9.82
QF-GPT	27.75	19.27	14.86	12.10	28.56	31.71	20.85
Prompt finetuning	32.79	23.69	18.65	15.39	37.55	40.81	72.55
Template-facet	38.56	26.32	20.87	17.61	33.87	46.02	100.0*
ZSFC + PPL	42.78	30.90	24.40	20.09	40.98	45.16	97.82
ZSFC + AutoScore	**43.42**	**31.95**	**25.61**	**21.61**	**41.03**	**47.87**	98.28
ZSFC + Cross	41.37	29.51	23.11	19.06	39.78	44.18	91.65
ZSFC + NTES	36.92	26.71	21.13	17.70	35.69	41.49	77.32
ZSFC + WSDM	41.80	29.69	23.03	18.88	38.52	44.19	**98.96**

Table 3: The effectiveness of question body only evaluation. The first number is the metric scores on question body. The second number (Δ) is the performance gap between question-body and full-question evaluation, which indicates the boosted portion by question templates.

Model	BLEU-1 (Δ)	BLEU-2 (Δ)	BLEU-3 (Δ)	BLEU-4 (Δ)	ROUGE-L (Δ)	METEOR (Δ)	Coverage (Δ)
Q-GPT-0	2.32 (-1.99)	1.37 (-1.12)	1.03 (-0.76)	0.87 (-0.58)	3.31 (-1.90)	3.71 (-1.89)	0.92 (-0.41)
QF-GPT-0	3.54 (-1.95)	2.14 (-1.07)	1.56 (-0.74)	1.27 (-0.57)	5.85 (-1.65)	6.95 (-1.37)	6.22 (-1.31)
Prompt-0	3.40 (-1.85)	2.07 (-0.93)	1.51 (-0.62)	1.23 (-0.47)	4.52 (-1.90)	5.58 (-1.80)	6.04 (-1.85)
Template-0	7.89 (-13.38)	5.22 (-10.04)	3.89 (-8.10)	3.21 (-6.75)	10.65 (-16.93)	11.13 (-12.91)	6.27 (-0.59)
Subject-constrained	14.51 (-14.65)	9.44 (-11.55)	6.97 (-9.55)	5.87 (-7.81)	14.51 (-14.08)	19.39 (-15.39)	9.45 (-0.37)
QF-GPT	16.38 (-11.37)	10.92 (-8.35)	8.17 (-6.69)	6.72 (-5.38)	18.41 (-10.15)	20.14 (-11.03)	20.21 (-0.64)
Prompt finetuning	24.13 (-8.66)	17.25 (-6.44)	13.41 (-5.24)	11.18 (-4.21)	31.84 (-5.71)	32.99 (-7.82)	71.41 (-1.14)
Template-facet	25.49 (-13.07)	15.62 (-10.7)	11.50 (-9.37)	9.47 (-8.14)	22.09 (-11.78)	38.58 (-7.44)	100.0 (0)*
ZSFC + PPL	36.80 (-5.98)	26.71 (-4.19)	20.85 (-3.35)	17.26 (-2.83)	**37.77** (-3.21)	40.58 (-4.58)	97.47 (-0.35)
ZSFC + AutoScore	34.88 (-8.54)	25.57 (-6.38)	20.53 (-5.08)	17.42 (-4.19)	33.88 (-7.15)	43.42 (-4.45)	95.83 (-2.45)
ZSFC + Cross	36.25 (-5.12)	26.57 (-2.94)	21.04 (-2.07)	17.59 (-1.47)	37.25 (-2.53)	40.45 (-3.73)	90.87 (-0.78)
ZSFC + NTES	29.46 (-7.46)	21.75 (-4.96)	17.56 (-3.57)	14.95 (-2.75)	27.97 (-7.72)	33.82 (-7.67)	77.32 (**0.00**)
ZSFC + WSDM	**38.51** (**-3.29**)	**28.34** (**-1.35**)	**22.50** (**-0.53**)	**18.99** (**+0.11**)	37.47 (**-1.05**)	**43.79** (**-0.40**)	**98.67** (-0.29)

We use ParlAI implementation[3] for cross-encoder, and the ConvAI3 winning team's implementation[4] for their ranker named NTES. We implement our own Weighted Sequential Dependency Model using the intuitively adjusted parameters: $\lambda_t = \lambda_o = \lambda_u = 1, \mu = 25$.

5 Results and Analyses

This section answers the research questions using our experiment results. In Table 2 and Table 3, we show the automatic metrics evaluation results respectively for the full-question evaluation and the question-body evaluation, as described in Section 4.3. In Table 4, we show the human annotation results of the compared models in the third research question.

Our first research question RQ1 is about the performance of existing methods for zero-shot clarifying question generation. The results are shown in the first four rows in the both tables. From the full question evaluation in table 2, we see that all these baselines struggle to produce any reasonable generations except for Template-0. However, we cannot conclude that Template-0 generates significantly better questions. Because when we compare the question body evaluation results in Table 3, we see the scores of Template-0 drop significantly. This means that its question body is not good, which implies that the reason for its higher score on full question evaluation is because of the question templates. In general, we find existing zero-shot GPT-2-based approaches cannot solve the clarifying question generation task effectively.

Our second research question RQ2 is about the effectiveness of facet information for facet-specific clarifying question generation. To answer this question, we compare our proposed zero-shot facet-constrained (ZSFC) methods with a facet-free variation of ZSFC named Subject-constrained which uses subject of the query as constraints. It would be unfair to compare the coverage metric

[3]https://github.com/facebookresearch/ParlAI
[4]https://github.com/ouwenjie03/Clariq_System
[4]The 100% coverage from the Template-facet method is not included in the coverage comparison.

Table 4: Human evaluations for models in RQ.3 according to major vote from 5 annotators. Our human evaluation results are aligned with our automatic evaluations. † and ‡ indicates $p < 0.05$ and $p < 0.0001$ statistical significance over other models.

model	Naturalness			Usefulness		
	Good	Fair	Bad	Good	Fair	Bad
QF-GPT	59.5%	14.4%	26.1%	11.5%	17.4%	71.1%
Prompt finetuning	43.7%	11.1%	45.2%	29.4%	22.4%	48.2%
Template-facet	57.4%	32.9%	9.6%	50.8%	36.0%	13.2%
ZSFC + WSDM	**82.6%**‡	13.9%	**3.5%**‡	**68.9%**‡	21.9%	**9.2%**†

between the two models because the Subject-constrained system does not access facet information. The gold references used for computing other metrics are the facet-specific clarifying questions from the dataset, thus it would also be incomprehensive to see the scores as the quality of clarifying question generation in general. Because the generated question could be reflecting another facet and get a low score on these metrics. However, these metrics could still be seen as the generation quality about these specific facets, which should intuitively be improved by adding facet as input, although not with naive GPT-2 [44].

From both the entire generation and question body evaluations, we see that all the ZSFC models significantly improve the Subject-constrained method across all the other evaluation metrics. The ZSFC models also drop less performance when switched to question-body evaluation, which suggests its better performance is more from the question body. In contrast to existing works, our study show that adequate use of facet information can significantly improve clarifying question generation quality.

The last research question RQ3 is whether our proposed zero-shot approach can perform the same or even better than existing facet-driven baselines. To answer this question, we compare our method with a simple clarifying question rewriting baseline and two finetuning baselines in the third section of the table. Among them, QF-GPT is the existing method [44], and Prompt finetuning is our proposed prompt-based finetuning method. We see that from both tables, our zero-shot facet-driven approaches are always better than the finetuning baselines. Our best-performing generation system, ZSFC+WSDM, improves the existing method QF-GPT by a large margin in the full-question evaluation and doubles its performance in question-body evaluation. When compared with the Template-facet baseline, ZSFC can outperform it in both tables. This implies these ZSFC system generations have more word overlaps from the reference beyond just the facet, which means the performance improvements are non-trivial. We also bold the smallest performance drop between the two evaluations. We can see that our ZSFC models have relatively minor performance drops, which means that they potentially generate better question bodies.

To validate the above conclusions, we employ human annotators to label the quality of generated questions. From Table 4, we can see that our proposed system ZSFC gives the best generation for both naturalness and usefulness. Specifically, it generates the most amount of good naturalness and usefulness questions and the least bad ones. This human evaluation result strengthens the conclusion from the automatic evaluation that our method is better than the baseline and supervised learning methods. We also notice that the Template-facet rewriting is a simple yet strong baseline that both finetuning-based methods are actually worse than it. However, ZSFC outperforms it by a large margin in both measures.

5.1 Ablation Study

We are also interested in whether question prompting is necessary for our system and which question ranker is the best in Section 3.2. To answer this question, we compare all five ranking methods mentioned in Section 3.2. For each (query, facet) pair in our dataset and each ranker, we will run the ranker on the eight question candidates using eight question templates and choose the top question from the ranked lists as the question generation of that ranker. The no-prompt-no-ranker approach will generate only one sentence with NeuroLogic Decoding with query as input and facet as constraints.

Here, we analyze the results of all the ranker variations in the last section from Table 2 and 3. The full-question evaluation results show that the AutoScore ranker performs best on all but the coverage metric. However, the question-body evaluation results suggest that the WSDM ranker performs the best. On the one hand, we believe the question-body evaluation results are more convincing by our previous analyses and examples of the two evaluation methods. On the other, we notice that when switched to question-body evaluation, the performances of AutoScore drop more than other methods, while WSDM almost does not change or even increase its scores. This could further suggest that part of the performances of AutoScore should be attributed to the question templates. We now explain why WSDM metrics have such low performance drops. Unlike other ranking methods, the way the WSDM scoring function is defined encourages generation to score higher with a high-quality question body since the question templates rarely contain facet or query subject words. Based on all the above observations and reasoning, we propose that WSDM is the best ranker.

6 Conclusion

In this work, we study the task of zero-shot clarifying question generation for conversational search. We propose to solve the task as a constrained language generation problem and present a concrete system. To demonstrate the power of our system, we answer three research questions, including comparing our zero-shot system with baseline and existing supervised learning approaches. All the experiment results have been evaluated using a variety of natural language generation metrics, and human evaluations are done for part of the results. The automatic metrics and human annotation results suggest our proposed zero-shot system outperforms the other compared approaches. Our work can be seen as both a solid zero-shot solution to the cold start problem of conversation search and a compelling demonstration of how large deep models benefit from properly integrating human knowledge.

References

[1] Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2020. ConvAI3: Generating clarifying questions for open-domain dialogue systems (ClariQ). *arXiv preprint arXiv:2009.11352* (2020).

[2] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*. 475–484.

[3] Avishek Anand, Lawrence Cavedon, Matthias Hagen, Hideo Joho, Mark Sanderson, and Benno Stein. 2021. Dagstuhl Seminar 19461 on Conversational Search: Seminar Goals and Working Group Outcomes. *SIGIR Forum* 54, 1, Article 3 (feb 2021), 11 pages. https://doi.org/10.1145/3451964.3451967

[4] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. Guided Open Vocabulary Image Captioning with Constrained Beam Search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 936–945. https://doi.org/10.18653/v1/D17-1098

[5] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics, Ann Arbor, Michigan, 65–72. https://aclanthology.org/W05-0909

[6] Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th conference of the european chapter of the association for computational linguistics*. 313–320.

[7] Michael Bendersky, Donald Metzler, and W. Bruce Croft. 2010. Learning Concept Importance Using a Weighted Dependence Model. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (New York, New York, USA) *(WSDM '10)*. Association for Computing Machinery, New York, NY, USA, 31–40. https://doi.org/10.1145/1718487.1718492

[8] Keping Bi, Qingyao Ai, and W Bruce Croft. 2021. Asking Clarifying Questions Based on Negative Feedback in Conversational Search. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. 157–166.

[9] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555* (2020).

[10] Charles L. A. Clarke, Nick Craswell, and Ian Soboroff. 2009. Overview of the TREC 2009 Web Track. In *TREC*.

[11] J Shane Culpepper, Fernando Diaz, and Mark D Smucker. 2018. Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (swirl 2018). In *ACM SIGIR Forum*, Vol. 52. ACM New York, NY, USA, 34–90.

[12] Van Dang and Bruce W Croft. 2010. Query reformulation using anchor text. In *Proceedings of the third ACM international conference on Web search and data mining*. 41–50.

[13] Kaustubh D Dhole. 2020. Resolving intent ambiguities by retrieving discriminative clarifying questions. *arXiv preprint arXiv:2008.07559* (2020).

[14] Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. *Can You Unpack That? Learning to Rewrite Questions-in-Context* (2019).

[15] Zuohui Fu, Yikun Xian, Yongfeng Zhang, and Yi Zhang. 2020. Tutorial on Conversational Recommendation Systems. In *Fourteenth ACM Conference on Recommender Systems*. 751–753.

[16] Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1371–1374.

[17] Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2022. Neural approaches to conversational information retrieval. *arXiv preprint arXiv:2201.05176* (2022).

[18] Claudia Hauff, Julia Kiseleva, Mark Sanderson, Hamed Zamani, and Yongfeng Zhang. 2021. Conversational Search and Recommendation: Introduction to the Special Issue.

[19] Chris Hokamp and Qun Liu. 2017. Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1535–1546. https://doi.org/10.18653/v1/P17-1141

[20] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969* (2019).

[21] Kimiya Keyvan and Jimmy Xiangji Huang. 2022. How to Approach Ambiguous Queries in Conversational Search? A Survey of Techniques, Approaches, Tools and Challenges. *ACM Computing Surveys (CSUR)* (2022).

[22] Antonios Minas Krasakis, Mohammad Aliannejadi, Nikos Voskarides, and Evangelos Kanoulas. 2020. Analysing the effect of clarifying questions on document ranking in conversational search. In *Proceedings of the 2020 acm sigir on international conference on theory of information retrieval*. 129–132.

[23] Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190* (2021).

[24] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 1823–1840. https://doi.org/10.18653/v1/2020.findings-emnlp.165

[25] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[26] Chang Liu, Jacek Gwizdka, Jingjing Liu, Tao Xu, and Nicholas J Belkin. 2010. Analysis and evaluation of query reformulations in different task types. *Proceedings of the American Society for Information Science and Technology* 47, 1 (2010), 1–9.

[27] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586* (2021).

[28] Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, et al. 2021. Neurologic a* esque decoding: Constrained text generation with lookahead heuristics. *arXiv preprint arXiv:2112.08726* (2021).

[29] Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Neurologic decoding:(un) supervised neural text generation with predicate logic constraints. *arXiv preprint arXiv:2010.12884* (2020).

[30] Gary Marchionini. 2006. Exploratory Search: From Finding to Understanding. *Commun. ACM* 49, 4 (apr 2006), 41–46. https://doi.org/10.1145/1121949.1121979

[31] Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. *arXiv preprint arXiv:1809.01984* (2018).

[32] Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*. Vol. 24. Elsevier, 109–165.

[33] Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. 2019. Cgmh: Constrained sentence generation by metropolis-hastings sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6834–6842.

[34] Wenjie Ou and Yue Lin. 2020. A clarifying question selection system from ntes_along in convai3 challenge. *arXiv preprint arXiv:2010.14202* (2020).

[35] 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*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[36] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[37] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*. 117–126.

[38] Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. *arXiv preprint arXiv:1805.04655* (2018).

[39] Sudha Rao and Hal Daumé III. 2019. Answer-based adversarial training for generating clarification questions. *arXiv preprint arXiv:1904.02281* (2019).

[40] Corbin Rosset, Chenyan Xiong, Xia Song, Daniel Campos, Nick Craswell, Saurabh Tiwary, and Paul Bennett. 2020. Leading conversational search by suggesting useful questions. In *Proceedings of The Web Conference 2020*. 1160–1170.

[41] Rodrygo LT Santos, Craig Macdonald, Iadh Ounis, et al. 2015. Search result diversification. *Foundations and Trends® in Information Retrieval* 9, 1 (2015), 1–90.

[42] Timo Schick and Hinrich Schütze. 2021. Few-shot text generation with natural language instructions. Association for Computational Linguistics.

[43] David Schlangen. 2004. Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*. 136–143.

[44] Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2021. Towards Facet-Driven Generation of Clarifying Questions for Conversational Search. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. 167–175.

[45] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *proceedings of the 24th ACM international on conference on information and knowledge management*. 553–562.

[46] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles LA Clarke. 2017. Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the 2017 chi conference extended abstracts on human factors in computing systems*. 2187–2193.