Towards Filling the Gap in Conversational Search: From Passage Retrieval to Conversational Response Generation

Weronika Łajewska
University of Stavanger
Stavanger, Norway
weronika.lajewska@uis.no

Krisztian Balog
University of Stavanger
Stavanger, Norway
krisztian.balog@uis.no

ABSTRACT

Research on conversational search has so far mostly focused on query rewriting and multi-stage passage retrieval. However, synthesizing the top retrieved passages into a complete, relevant, and concise response is still an open challenge. Having snippet-level annotations of relevant passages would enable both (1) the training of response generation models that are able to ground answers in actual statements and (2) automatic evaluation of the generated responses in terms of completeness. In this paper, we address the problem of collecting high-quality snippet-level answer annotations for two of the TREC Conversational Assistance track datasets. To ensure quality, we first perform a preliminary annotation study, employing different task designs, crowdsourcing platforms, and workers with different qualifications. Based on the outcomes of this study, we refine our annotation protocol before proceeding with the full-scale data collection to gather annotations for 1.8k question-paragraph pairs. The process of collecting data at this magnitude also led to multiple insights about the problem that can inform the design of future response-generation methods.

CCS CONCEPTS

• **Information systems** → **Presentation of retrieval results**; **Crowdsourcing**; **Information extraction**.

KEYWORDS

Conversational search, Conversational response generation, Snippet annotation, Crowdsourcing

ACM Reference Format:
Weronika Łajewska and Krisztian Balog. 2023. Towards Filling the Gap in Conversational Search: From Passage Retrieval to Conversational Response Generation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3583780.3615132

1 INTRODUCTION

A large fraction of research on conversational information seeking (CIS) to date has focused on the problem of retrieving relevant passages. The Conversational Assistance track at the Text Retrieval Conference (TREC CAsT) [6–8, 21] has played a major role in enabling research on this task by developing a series of reusable test collections. The task of conversational passage retrieval requires advances in query rewriting [18, 31, 32] and can also directly benefit from research on multi-stage passage retrieval [20]. However, identifying relevant passages is only an intermediate step. Ultimately, the information contained in these passages would need to be synthesized into a single answer. *Conversational response generation* is the task of encapsulating the most relevant pieces of information in an easily consumable unit [5]. Including it in the CIS pipeline would increase the naturalness of the conversation [28, 29].

There are at least two main challenges involved in the task of response generation: identifying key pieces of information from relevant results (e.g., paragraphs) and summarizing them in a concise answer. Correspondingly, Ren et al. [26] propose to split the task into two stages: (1) identification of supporting snippets and (2) summarization of selected snippets. In this paper, we focus on the problem of (1), and more specifically on building a snippet dataset with high-quality annotations using crowdsourcing.

The significance of being able to identify relevant snippets is twofold. First, it enables the training of models that can ground the generated answers in actual statements. Natural language generation models are susceptible to hallucinations, especially if the query is insufficiently covered in the corpus, or the retrieved documents contain redundant, complementary, or contradictory information [15]. Therefore, employing abstractive summarization methods on top of relevant snippets identified can help to mitigate this problem and provide more control over the generation process, much in the spirit of the two-step process proposed in [26]. Second, it would enable automatic evaluation of the generated responses quantitatively, in terms of relevant information nuggets included [23]. Response summarization in CIS systems has been piloted in the most recent edition of TREC CAsT [21], where the quality of answer summaries is evaluated by human judges along three dimensions: relevance, naturalness, and conciseness [21]. Having annotations of relevant snippets would enable automatic evaluation of answers in terms of completeness.

Even though crowdsourcing has become an established means of collecting human annotations at scale, ensuring data quality can be challenging [9]. Indeed, we demonstrate that the seemingly straightforward task of highlighting relevant snippets may not be so simple and deserves more close attention.

In this paper, we first investigate what are effective task designs and trade-offs between worker qualifications and costs to perform the task of snippet annotations. Specifically, we consider paragraph-

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3615132

and sentence-level snippet annotation interfaces, multiple crowdsourcing platforms, and crowd workers with different qualifications as well as expert annotators. Measuring the quality of annotations is challenging because relevant snippet selection is subjective and often there are multiple correct sets of snippets in a given passage. We evaluate the resulting annotations in terms of inter-annotator agreement and similarity to expert annotations using text similarity measures adapted to this task.

Based on the results of our preliminary study, we set out to create a large-scale dataset, CAsT-snippets, which enriches the TREC CAsT 2020 and 2022 datasets with snippet-level answer annotations. We follow a setup in which we closely work with a selected pool of highly engaged crowd workers in order to ensure high data quality. Our findings from this data collection effort reveal numerous associated challenges that can help inform the design of response generation methods in future work.

The resources developed in this study (annotated data and code for computing evaluation measures) are made publicly available at https://github.com/iai-group/CAsT-snippets. An extended version of this paper is available on arXiv.

2 RELATED WORK

Research on conversational response generation has attracted a lot of attention in task-oriented dialogue systems [2, 19, 24], question answering [1], open-domain chatbots [10, 33], and most recently in conversational information seeking, as part of TREC CAsT'22 [21]. The performance of response generation is commonly evaluated using automatic similarity measures for natural language generation tasks such as BLEU [16, 22], and ROUGE [17]. However, some dimensions are not reliably covered by currently available automatic metrics and require manual evaluation (e.g., coherence and relevance) [12], while others (e.g., completeness) can be evaluated automatically, provided that more fine-grained annotations are available. Information nuggets, defined as *minimal, atomic units of relevant information* of retrieved documents, have been proposed as an alternative to automatically assign relevance judgments to documents and/or evaluate retrieval systems [23]. Our work aims to contribute to this type of evaluation by studying ways to collect snippet-level annotations. A task similar to snippet annotation (or information nuggets identification) has been broadly researched in QA systems. In most available datasets for reading comprehension focused mainly on factoid questions, the generated response is a single entity or a short segment of text from the passage [3, 4, 25, 27].

Crowdsourcing provides a scalable means to the completion of large amounts of labeling or annotation tasks that require human intelligence [13]. The actual quality of the results is influenced by the workers, software platform [30], task design [11], and quality measures employed [9]. This paper attempts to understand what setup is needed to effectively perform the task of snippet annotation.

Relevant annotations efforts include QuaC [4], which is dataset of QA dialogues. However, it is limited to sections of Wikipedia articles and contains only dialogues about a biased sample of entities of type person. Queries in CAsT datasets are much more diverse, both in terms of the expected type of answer and in the topics discussed. Most relevant to our paper is the work by Ren et al. [26], where crowd workers are asked to respond to queries from the TREC CAsT'19 dataset while being presented with SERPs. The response generation task is divided into three stages: (optional) query rewriting, finding supporting sentences in results displayed on a SERP, and summarizing them into a short conversational response. We focus only on the supporting evidence finding step, which is performed on a finer (snippet-level) granularity, and explore various task designs to ensure high data quality.

3 DATASET

We perform annotations on the TREC CAsT 2020 and 2022 datasets.[1] Each dataset comprises of set of information-seeking dialogues (i.e., topics) with a sequence of questions (i.e., queries) within each. The input to the snippet annotation task consists of queries and corresponding passages. We consider the top 5 passages for each query with respect to their relevance labels in the ground truth (ranging from 0 to 4). If there are fewer than 5 passages available for the query at the highest relevance level, then we fill up the remaining slots with passages one relevance level below. If there are more passages available, then we cluster them using k-means clustering and pick a random passage per cluster. For example, if we have 3 highly relevant passages for a given query and 10 relevant passages, we choose all the passages with relevance level 4 and populate the remaining two places by splitting the passages with a relevance level 3 into two clusters and then choosing a random passage from each cluster. Selecting the passages for annotation this way ensures that they are both relevant and diverse. Even though we mostly consider highly relevant and relevant passages, some of them do not contain a direct answer to the question, which makes the snippet annotation task even more challenging.

4 PRELIMINARY STUDY

To ensure that we get high-quality snippet-level annotations, we first perform a preliminary study where we compare different different task designs, platforms, and worker pools, by annotating two topics selected from the TREC CAsT'22 dataset, with markedly different characteristics, comprising of 22 queries in total.

4.1 Task Designs

We task crowd workers with the identification of snippets in a provided text that contains key pieces of the answer to a given query. Text snippets are required to be short, concise, informative, self-contained, and cannot overlap. Each snippet is supposed to contain one piece of information, so it can be treated as an information nugget. Specifically, we identify snippets in paragraphs that have been labeled as relevant answers to the question. These passages can be long, which makes the annotation task cognitively demanding. Therefore, we consider two designs of the task: paragraph-based and sentence-based; see Figure 1.

In the *paragraph-based* annotation task, workers are asked to identify all text snippets in a given passage that are relevant to the input query. Since paragraphs can be lengthy, we also consider a simplified, *sentence-based* variant of this task, which lets workers operate on the significantly shorter text and enforces shorter text snippet selection. Specifically, the task is divided into: (1) relevant

[1]The 2019 dataset has relatively low complexity compared to these two, while the 2021 dataset provides relevance assessments on the level of documents instead of passages.

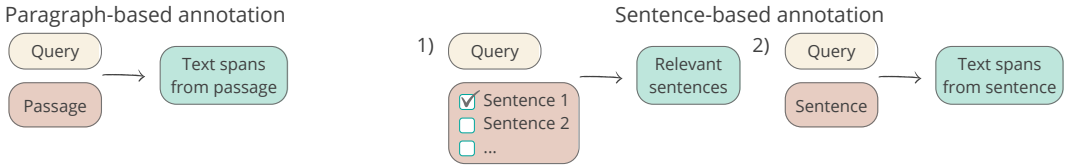

Figure 1: Illustration of different designs for the snippet annotation task.

sentence selection, and (2) snippet annotation in relevant sentences. In sub-task (1), crowd workers are presented with a question and a passage that is split into sentences. They are asked to choose sentences that contain information relevant to the query. This is a straightforward task that resembles extractive summarization [34]. Sub-task (2) is very similar to the paragraph-based annotation task, the only difference is that workers are presented with a relevant sentence instead of an entire passage.

4.2 Platforms and Workers

We set up the annotation task on two crowdsourcing platforms: Amazon MTurk and Prolific. MTurk offers an easily customizable web-based annotation interface and it is possible to filter workers based on qualifications. Prolific has more limited options in terms of the annotation interface, but the qualification of workers is claimed to be higher than on MTurk.[2] Additionally, we employ a group of expert annotators (Ph.D. students) who have been trained to perform this annotation task; they also use the MTurk platform, but in sandbox mode, i.e., without receiving payment. The paragraph-based annotation task, which is regarded as the cognitively more demanding variant, is performed with workers from both crowdsourcing platforms as well as with expert annotators. The sentence-based variant of the task is executed only on MTurk. All tasks on MTurk are performed with both regular and master workers.[3]

4.3 Evaluation Measures

Traditional metrics for inter-annotator agreement such as Fleiss' Kappa or Krippendorff's Alpha are designed to assess categorical annotations and rely on a binary notion of agreement. In our case, we are more interested in measuring the degree to which snippets selected by different workers overlap. We define inter-annotator agreement in terms of Jaccard similarity. Given an input text t annotated by n workers, we count the length of the snippets chosen by all annotators and divide it by the total length of snippets chosen by any annotator. The intersection and union of snippet intervals is calculated on the character level. We also consider a less strict variant of the measure, termed Jaccard$_k$ (J_k), which takes only those intervals into account that are chosen by at least k annotators.

To measure the similarity of snippet annotations by crowd workers against reference annotations by experts, we follow a logic similar to ROUGE-1, which considers the overlap of unigrams between the system and reference summaries [17]. Specifically, we employ the ROUGE-like measures proposed in [14]. For every input text t, we have annotations made by n different crowd workers (w_i) and reference annotations by m different experts (e_j). First, we define precision (recall) of the snippets in text t between a pair of annotators w_i and e_j as the length of the snippets chosen by both the crowd worker and expert annotator, and divide it by the length of snippets chosen by crowd worker (expert annotator). We compute the F1 score as the harmonic mean of precision and recall. Next, we average these measures for a given crowd worker i against all (m) expert annotations by taking the mean. Finally, we aggregate the annotations across all (n) crowd workers by averaging precision, recall, and F1 against all expert annotations over all crowd workers.

[2]https://www.prolific.co/prolific-vs-mturk
[3]MTurk Master is a qualification earned through a proven track record of quality work.

Table 1: Inter-annotator agreements (J and J$_k$) and similarity against reference (expert) annotations ($\overline{F1}$). The number of annotators for every input text is shown in parentheses.

Task Variant	Annotator	J	J$_k$			$\overline{F1}$
			$k=4$	$k=3$	$k=2$	
Paragraph	MTurk regular ($n=5$)	0.02	0.08	0.21	0.48	0.36
	MTurk master ($n=5$)	0.18	0.35	0.53	0.73	**0.54**
	Prolific ($n=5$)	0.14	0.27	0.44	0.65	0.50
	Expert ($m=3$)	0.25	-	-	0.54	-
Sentence	MTurk regular ($n=3$)	0.35	-	-	0.71	0.31
	MTurk master ($n=3$)	0.47	-	-	0.76	0.41

4.4 Results

We report on the inter-annotator agreement and similarity against reference annotations for two topics selected for this preliminary study in Table 1. (Payments and average task completion times are reported in the extended version.) The total cost was $1.2k.

In the paragraph-based variant, we observe better agreement (J) between MTurk masters than between Prolific workers, yet there is a big gap between crowd workers and experts. The relative ordering is: MTurk masters > Prolific > MTurk regular, which also holds for the more relaxed version of the measure (J_k). We notice that for J_2, the agreement between expert annotators is lower than for MTurk masters and Prolific workers; however, there are only 3 experts (vs. 5 crowd workers), hence it is not fair to directly compare these numbers. The generally low agreement scores highlight the difficulty of the task in the paragraph-based form.

On the the simplified sentence-based variant, we indeed observe a much higher agreement between MTurk workers.[4] Also, the differences between regular workers and masters are not as large as in the paragraph-based variant. We note that the two task variants (sentence-based and paragraph-based) cannot be compared directly in terms of inter-annotator agreement because the probability of choosing the same snippets by different workers is much higher in a single sentence than in an entire paragraph. Overall, the similarity with experts is higher in case of paragraph-level annotations than for sentence-level annotations.

[4]Given that MTurk masters outperformed Prolific workers in the paragraph-based variant, sentence-based annotations are only performed on MTurk.

4.5 Discussion

Our preliminary exploration of different task designs, platforms, and workers has led us to the conclusion that the highest-quality annotations for this specific task can be collected on the MTurk platform using a paragraph-based task design. The main challenge in collecting snippet annotations turned out to be the process of quality control that cannot be automated due to the nature of this task. Even for expert annotators, who performed the task attentively, the inter-annotator agreement is low. Therefore, a low similarity between snippets selected by a worker and reference annotations does not imply that the worker did an inferior job. Moving forward to collecting annotations at scale, we opt for recruiting a smaller group of crowd workers, using a qualification task, and working closely with them by providing continuous feedback on their work.

5 DATA COLLECTION

This section describes our large-scale data collection effort. For each of the 371 queries in the TREC CAsT 2020 and 2022 datasets, the top 5 passages are annotated by 3 crowd workers, resulting in a total of 1,855 query-passage pairs.

5.1 Setup

The annotation task was released only to a small group of trained crowd workers, who were selected through a qualification task. The qualification task contained a detailed description of the problem at hand, examples of correct annotations, a quiz, and 10 query-passage pairs to be annotated; it was made available to both master and regular MTurk workers to reach a bigger audience. From the 20 workers that completed the qualification task, we chose 15 that had the highest quality results (independently of their MTurk Master qualification). Each worker received feedback on the provided responses and was given an opportunity to ask their own questions about the task. Several rounds of discussion that emerged from the qualification task resulted in an extended set of guidelines addressing the challenging aspects of the annotation task. The extended guidelines are made available in the online repository.

The process of data collection was divided into daily batches and conducted over a period of approx. two weeks. The reason was to both avoid worker fatigue and also to allow for continuous feedback along the way. Each batch contained questions about one specific topic, which amounts to 46 query-passage pairs on average, and was annotated by 3 different workers. Workers received $0.3 for each query-passage pair. A bonus of $2 was paid for every batch completed within 24 hours upon release. The total cost was $2.1k.

The training of the annotators did not end at the qualification task, but continued throughout the whole data collection process. Crowd workers were provided with feedback after each submitted batch. From each batch, random data samples with low agreement were selected and verified manually by an expert (the main author of the paper). We used Slack as the main communication platform; there, workers could also share challenging cases and benefit collectively from discussions and from expert guidance.

5.2 Statistics

In comparison to the results of the preliminary study (cf. Table 1) on the same set of queries, we find that the inter-annotator agreement (J=0.38 and J_2 =0.62) exceeds even that of expert annotations, and the similarity with expert annotations ($\overline{F1}$ =0.54) matches those of the best-performing MTurk master workers. These results indicate that the collected data is of high quality and attest to the success of our annotation setup with continuous feedback.

Table 2: Comparison against other datasets.

Dataset	Input text	Avg. snippet length (tokens)	#snippets per annotation
CAsT-snippets	Paragraph	39.6	2.3
SaaC [26]	Top 10 passages	23.8	1.5
QuaC [4]	Wikipedia article	14.6	1

Table 2 provides a comparison against other related datasets. We note that there are not only more snippets annotated for each input text in our dataset, but they are also longer on average, which follows from the information-seeking nature of queries.

We note that there is a number of query-passage pairs where annotators did not find any snippet relevant to the query, despite the passage being labeled as relevant by TREC assessors (77 such passages selected by all three annotators and 111 selected by two of the annotators).

6 CONCLUSIONS

We have introduced CAsT-snippets, a high-quality dataset for conversational information seeking containing snippet-level annotations for all queries in the TREC CAsT 2020 and 2022 datasets. Our annotation effort was informed by a preliminary study, where we explored various task designs, platforms, and workers pools. Based on the results, we opted for a setup where we closely worked with a pool of highly engaged crowd workers, releasing tasks in daily batches and providing continuous feedback.

Our direct communication with crowd workers throughout the data annotation process revealed multiple challenges that need to be addressed in conversational response generation: (1) Selecting spans for questions when only a partial answer is present is challenging and appears to be highly subjective. (2) Temporal considerations may exclude some spans as they are not valid answers given the time specified in the query. However, assessing the temporal validity of text may be challenging based solely on short text passages without a larger context. (3) Passages originating from blogs or comments very often contain subjective opinions. Should such subjective opinions be marked up as answers? (4) What kind of background knowledge should be assumed when the passage does not contain a direct answer but the answer may be inferred from the text? (5) How much content is needed for open-ended questions? (6) When is evidence or additional information needed for a factoid question and when is an entity alone sufficient as an answer?

Our dataset enables the development of answer generation methods that are grounded in relevant snippets in paragraphs as well as allows for the automatic evaluation of the generated answers in terms of completeness; a training/test split is provided for such use.

ACKNOWLEDGMENTS

This research was supported by the Norwegian Research Center for AI Innovation, NorwAI (Research Council of Norway, project number 309834).

REFERENCES

[1] Ashutosh Baheti, Alan Ritter, and Kevin Small. 2020. Fluent Response Generation for Conversational Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL '20)*. 191–207.

[2] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Findings of the Association for Computational Linguistics: EMNLP 2018 (EMNLP '18)*. 5016–5026.

[3] Daniel Fernando Campos, Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, Li Deng, and Bhaskar Mitra. 2016. MS MARCO: A Human Generated MAchine Reading Comprehension Dataset. arXiv:1611.09268 [cs.CL]

[4] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Findings of the Association for Computational Linguistics: EMNLP 20 (EMNLP '18)*. 2174–2184.

[5] 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). *ACM SIGIR Forum* 52, 1 (2018), 34–90.

[6] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2019. TREC CAsT 2019: The Conversational Assistance Track Overview. In *The Twenty-Eighth Text REtrieval Conference Proceedings (TREC '19)*.

[7] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. TREC CAsT 2020: The Conversational Assistance Track Overview. In *The Twenty-Ninth Text REtrieval Conference Proceedings (TREC '20)*.

[8] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2021. TREC CAsT 2021: The Conversational Assistance Track Overview. In *The Thirtieth Text REtrieval Conference Proceedings (TREC '21)*.

[9] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality Control in Crowdsourcing. *Comput. Surveys* 51, 1 (2018), 1–40.

[10] Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar R Zaiane. 2019. Augmenting Neural Response Generation with Context-Aware Topical Attention. In *Proceedings of the First Workshop on NLP for Conversational AI (ACL '19)*. 18–31.

[11] Carsten Eickhoff. 2018. Cognitive Biases in Crowdsourcing. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM '18)*. 162–170.

[12] A. R. Fabbri, Wojciech Kryscinski, Bryan McCann, Richard Socher, and Dragomir R. Radev. 2020. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics* 9 (2020), 391–409.

[13] Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze. 2015. Human Beyond the Machine: Challenges and Opportunities of Microtask Crowdsourcing. *IEEE Intelligent Systems* 30 (2015), 81–85.

[14] Neslihan Iskender, Robin Schaefer, Tim Polzehl, and Sebastian Möller. 2021. Argument Mining in Tweets: Comparing Crowd and Expert Annotations for Automated Claim and Evidence Detection. In *International Conference on Applications of Natural Language to Data Bases (NLDB '21)*. 275–288.

[15] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of Hallucination in Natural Language Generation. arXiv:2202.03629 [cs.CL]

[16] Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2017. The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association for Computational Linguistics* 6 (2017), 317–328.

[17] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of Workshop on Text Summarization Branches Out on Annual Meeting of the Association for Computational Linguistics (ACL '04)*. 74–81.

[18] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2021. Multi-stage Conversational Passage Retrieval: An Approach to Fusing Term Importance Estimation and Neural Query Rewriting. *ACM Transactions on Information Systems* 39, 4 (2021), 1–29.

[19] Phillip Lippe, Pengjie Ren, Hinda Haned, Bart Voorn, and M. de Rijke. 2020. Diversifying Task-oriented Dialogue Response Generation with Prototype Guided Paraphrasing. arXiv:2008.03391 [cs.CL]

[20] Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, Dense, and Attentional Representations for Text Retrieval. *Transactions of the Association for Computational Linguistics* 9 (2021), 329–345.

[21] Paul Owoicho, Jeffrey Dalton, Mohammad Aliannejad, Leif Azzopardi, Johanne R. Trippas, and Svitlana Vakulenko. 2022. TREC CAsT 2022: Going Beyond User Ask and System Retrieve with Initiative and Response Generation. In *The Thirty-First Text REtrieval Conference Proceedings (TREC '22)*.

[22] 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 (ACL '02)*. 311–318.

[23] Virgil Pavlu, Shahzad Rajput, Peter B. Golbus, and Javed A. Aslam. 2012. IR System Evaluation Using Nugget-based Test Collections. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM '12)*. 393–402.

[24] Jiahuan Pei, Pengjie Ren, Christof Monz, and M. de Rijke. 2019. Retrospective and Prospective Mixture-of-Generators for Task-oriented Dialogue Response Generation. In *European Conference on Artificial Intelligence (ECAI '19)*.

[25] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Findings of the Association for Computational Linguistics: EMNLP 2016 (EMNLP '16)*. 2383–2392.

[26] Pengjie Ren, Zhumin Chen, Zhaochun Ren, E. Kanoulas, Christof Monz, and M. de Rijke. 2021. Conversations with Search Engines: SERP-based Conversational Response Generation. *ACM Transactions on Information Systems* 39, 4 (2021), 1–29.

[27] Chuanqi Tan, Furu Wei, Nan Yang, Weifeng Lv, and M. Zhou. 2017. S-Net: From Answer Extraction to Answer Generation for Machine Reading Comprehension. arXiv:1706.04815 [cs.CL]

[28] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the Design of Spoken Conversational Search: Perspective Paper. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. 32–41.

[29] Johanne R. Trippas, Damiano Spina, Paul Thomas, Mark Sanderson, Hideo Joho, and Lawrence Cavedon. 2019. Towards a Model for Spoken Conversational Search. *Information Processing & Management* 57, 2 (2019), 102–162.

[30] Donna Vakharia and Matthew Lease. 2013. Beyond AMT: An Analysis of Crowd Work Platforms. arXiv:1310.1672 [cs.CY]

[31] Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question Rewriting for Conversational Question Answering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM '21)*. 355–363.

[32] Svitlana Vakulenko, Nikos Voskarides, Zhucheng Tu, and Shayne Longpre. 2021. A Comparison of Question Rewriting Methods for Conversational Passage Retrieval. In *European Conference on Information Retrieval (ECIR '21)*. 418–424.

[33] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, M. Zhou, and Wei-Ying Ma. 2016. Topic Aware Neural Response Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI '16)*. 3351–3357.

[34] Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, M. Zhou, and Tiejun Zhao. 2018. Neural Document Summarization by Jointly Learning to Score and Select Sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL '18)*. 654–663.