A Dataset for Joint Conversational Search and Recommendation[⋆]

Marco Alessio[1,4,*,†], Simone Merlo[2,*,†], Tommaso Di Noia[3], Guglielmo Faggioli[2], Marco Ferrante[5], Nicola Ferro[2], Cristina Ioana Muntean[1], Franco Maria Nardini[1], Fedelucio Narducci[3], Raffaele Perego[1], Giuseppe Santucci[6] and Nicola Viterbo[3]

[1] *Institute of Information Science and Technologies at National Research Council of Italy, Via G. Moruzzi 1, 56124 Pisa, Italy*

[2] *Department of Information Engineering at University of Padua, Via G. Gradenigo, 6/b, 35131 Padua, Italy*

[3] *Department of Electrical and Information Engineering at Politecnico di Bari, Via E. Orabona, 4, 70125 Bari, Italy*

[4] *Department of Computer Science at University of Pisa, Largo B. Pontecorvo, 3, 56127 Pisa, Italy*

[5] *Department of Mathematics at University of Padua, Via Trieste, 63, 35121 Padua, Italy*

[6] *Department of Computer, Control, and Management Engineering at Sapienza University of Rome, Piazzale A. Moro, 5, 00185 Rome, Italy*

Abstract

Conversational Information Access systems have undergone widespread adoption due to the natural and seamless interactions they enable with the user. In particular, they provide an effective interaction interface for both Conversational Search (CS) and Conversational Recommendation (CR) scenarios. Despite their inherent similarities, current research frequently address CS and CR systems as distinct and isolated entities. The integration of these two capabilities would enable to address complex information access scenarios, including the exploration of unfamiliar features of recommended products, which leads to richer dialogues and enhanced user satisfaction. At current time, the evaluation of integrated by-design CS and CR systems is severely hindered by the limited availability of comprehensive datasets that jointly address both tasks. To bridge this gap, we introduce CoSRec[1], the first dataset for joint Conversational Search and Recommendation (CSR) evaluation. The CoSRec test set includes 20 high-quality conversations, with human-made annotations for the quality of conversations, and manually crafted relevance judgments for products and documents. In addition, we provide auxiliary training resources, including partially annotated dialogues and raw conversations, to support diverse learning paradigms. CoSRec is the first resource to model CS and CR tasks within a unified framework, facilitating the design, development, and evaluation of systems capable of dynamically alternating between answering user queries and offering personalized recommendations.

Keywords

Conversational Search, Conversational Recommendation, Joint Information Retrieval and Recommendation

1. Introduction

Conversational Agents (CAs) had a major impact on information access by enabling natural interaction. However, CAs introduce some additional challenges since they must handle dynamic and complex natural language conversations. Information Retrieval (IR) and Recommender Systems (RS) represent

[1] The dataset and code are available at: https://github.com/CAMEO-22/CoSRec

IIR2025: 15th Italian Information Retrieval Workshop, 3th - 5th September 2025, Cagliari, Italy

[⋆] This work is an extended abstract of [1].

[*] Corresponding author.

[†] These authors contributed equally.

marco.alessio@isti.cnr.it (M. Alessio); simone.merlo@phd.unipd.it (S. Merlo); tommaso.dinoia@poliba.it (T. D. Noia); guglielmo.faggioli@unipd.it (G. Faggioli); ferrante@math.unipd.it (M. Ferrante); ferro@dei.unipd.it (N. Ferro); cristina.muntean@isti.cnr.it (C. I. Muntean); francomaria.nardini@isti.cnr.it (F. M. Nardini); fedelucio.narducci@poliba.it (F. Narducci); raffaele.perego@isti.cnr.it (R. Perego); santucci@dis.uniroma1.it (G. Santucci); n.viterbo@studenti.poliba.it (N. Viterbo)

0009-0008-5043-2174 (M. Alessio); 0009-0003-8003-4795 (S. Merlo); 0000-0002-0939-5462 (T. D. Noia); 0000-0002-5070-2049 (G. Faggioli); 0000-0002-0894-4175 (M. Ferrante); 0000-0001-9219-6239 (N. Ferro); 0000-0001-5265-1831 (C. I. Muntean); 0000-0003-3183-334X (F. M. Nardini); 0000-0002-9255-3256 (F. Narducci); 0000-0001-7189-4724 (R. Perego); 0000-0003-4350-1123 (G. Santucci); 0009-0007-6723-0969 (N. Viterbo)

the information access systems that benefit most from conversational interfaces. Conversational Search (CS) systems assist users in refining their information needs through multi-turn dialogues, while Conversational Recommendation (CR) systems guide users in exploring a catalogue of items to identify optimal recommendations. CS and CR share significant commonalities, as both rely on iterative, multi-turn interactions to progressively refine user needs [2] despite having different goals. The development of Conversational Search and Recommendation (CSR) systems, which support both search and recommendation, could improve the user satisfaction. Indeed, when seeking for a recommendation, it is common to look for additional information about the recommended items (and the other way around). Recent studies in the joint IR and RS field [3, 4, 5, 6, 7], though not yet conversational, have demonstrated the benefits of modeling these tasks together. This suggests that integrating CS and CR into a unified conversational framework could lead to similar improvements. Historically, CS and CR have been treated in isolation. This approach has hindered the development of joint conversational search and recommendation systems. The major obstacle towards the development of joint CSR systems is the lack of publicly available resources suitable for training and evaluation. While rich datasets exist for individual tasks, *e.g.,* the TREC CAsT collections [8, 9, 10, 11] for search and REDIAL [12] for recommendation, there is a notable absence of datasets tailored for joint scenarios.

To facilitate the development of CSR systems, we introduce and release CoSRec, the first large-scale dataset explicitly designed for joint CSR tasks. CoSRec comprises approximately 9,000 user-system conversations generated by a Large Language Model (LLM) in the product search and recommendation domain. These conversations encompass a variety of interactions, including pure search, pure recommendation, and mixed search-and-recommendation utterances. As a result, a CSR system tested on CoSRec must accurately interpret the user's intent in each utterance and respond appropriately, taking into account the context of previous interactions. To ensure the quality of the dataset, a sample of approximately 3% of the conversations has been manually annotated to identify user intents and assess overall quality. Additionally, for 20 high-quality conversations, we provide utterance-level human-generated relevance judgments for items or documents, depending on the intent of the utterance. These annotations enable precise and effective evaluation of joint CSR systems.

Our contributions can be summarized as follows: (1) *Release of* `CoSRec-Raw`: a dataset comprising approximately 9,000 automatically generated conversations for joint search and recommendation tasks. Alongside the dataset, we provide a toolkit to generate additional conversations, enabling further research and scalability. (2) *Release of* `CoSRec-Crowd`: a subset of over 290 conversations manually annotated for quality. Each utterance in these conversations is labeled with its intent (search, recommendation, or joint search and recommendation), offering valuable insights for intent recognition and system evaluation. (3) *Release of* `CoSRec-Curated`: a high-quality subset of 20 deeply annotated conversations. For each utterance, we include manual (personalized) annotations identifying relevant passages or items, enabling precise and granular evaluation of CSR systems.

2. The Structure of CoSRec

CoSRec is a novel multi-domain conversational dataset designed to jointly address CS and CR by leveraging product-related dialogues as a natural application domain. Following the classic Cranfield paradigm for offline evaluation, CoSRec includes three elements: a set of *information needs, i.e.,* conversations, a *document corpus and item catalogue*, and a set of human-made *annotations*. At the same time, these elements have been adapted to fit our CSR scenario.

2.1. Information Needs and Corpora

In the CoSRec dataset, information needs are represented by conversations. CoSRec includes 9,249 conversations split into 3 partitions: `CoSRec-Raw`: 8,938 non-annotated conversations containing 71,656 utterances; `CoSRec-Crowd`: 291 human-annotated conversations including 2,329 utterances; `CoSRec-Curated`: 20 deeply human-annotated conversations containing 150 utterances. Each conversation is a multi-turn dialogue between a user and a system, where each turn corresponds to a user's utterance and a system's response. Hence, each user's utterance represents one or more information needs the system must satisfy, among: (1) **Search:** the user asked for general information about a topic related to the product they are discussing; (2) **Recommendation:** the user asks for some products to be suggested, according to her requirements; (3) **Product Detail:** the user inquires about details of the product being discussed (*e.g.,* price, brand, size). The type of answer is the main difference between "search" and "product detail". A product detail question can be answered by inspecting the product's description. On the other hand, search intents denote open-ended questions whose answers are likely to be found on an external corpus.

These information needs require the system to answer with items drawn from a catalogue, *i.e.,* recommendation and product detail intents, or with information retrieved from a corpus of documents, *i.e.,* search intent. Therefore, we need a corpus and a catalogue to serve as the foundation for the system's answers during evaluation. To this end, we rely on two publicly available resources: MS-MARCO v2.1 [13] comprising over 113.5M passages for search intents and Amazon Reviews [14] with 12.3M items for recommendation intents. Each intent is associated with a "canonical formulation" describing the information need in isolation and a series of human-made reformulations. Every conversation in CoSRec is associated with at least 3 user profiles. Such user profiles are composed of two elements: a brief textual summary of the user's interests and a set of keywords, constructed using the text of the users' past reviews. Hence, they can be used to personalize the CSR system's responses.

2.2. Human Annotations: Conversations Quality Assessment and Intents Labeling

Among the 9,249 conversations included in CoSRec, a subset of 311 (∼3%) are manually annotated to assess their quality. In particular, our annotation process involved 99 semi-expert human annotators. Each conversation was assigned to five annotators to ensure that at least three quality assessments were available for each conversation. Such quality assessments are given on a 1 to 5 scale and concern 4 aspects [15]: (1) **Fluency:** a conversation is fluent when it is well organized, in regular English grammar, easy to understand, and has a continuous flow; (2) **Informativeness:** a conversation is informative when the utterances include substantial content, communicate the user's needs, or deliver valuable information; (3) **Logicality (*a.k.a.,* Inverse Perplexity):** a conversation has a high logicality when its utterances are organized according to a logical flow and align with common reasoning; (4) **Coherence:** a conversation is coherent when the user and the system follow each other without unexpected or inappropriate utterances. Furthermore, given the specific product search setting, the user's final utterance must be consistent with the needs expressed during the dialogue.

The human annotators also associated intent labels and stand-alone formulations to each utterance of the conversations. Every utterance is annotated with zero, one, or more among "search", "recommendation", and "product detail" intent labels. Additionally, for each intent, the annotator provides a self-explanatory textual description of the information need, independent of the conversation's context, as it fully encapsulates it. Based on the quality assessment results, 20 high-quality conversations are then selected and refined to form the `CoSRec-Curated` dataset, while the remaining 291 form the `CoSRec-Crowd` partition. These annotations are released as they are for the `CoSRec-Crowd` partition of the dataset. In contrast, for the 20 `CoSRec-Curated` conversations, the authors of this paper further refined the labels by reviewing cases where annotators did not reach unanimity. Through discussion, they assigned the most appropriate label. As with intent labeling, the stand-alone formulations were carefully reviewed to correct typographic errors and ensure consistency.

2.3. Human Annotations: Relevance Judgments

The `CoSRec-Curated` portion of the dataset contains a total of 17,464 relevance judgments for user intents related to search and recommendation. Each intent has between 26 and 452 judgments, with an average of 166.3. The query-document (search) or query-product (recommendation) pairs to be assessed have been selected by retrieving, for each query, 1000 documents or products with BM25, by re-ranking them using SPLADE [16], TCT ColBERT [17] and Contriever [18] and by pooling the re-ranked results with a pooling depth of 10. During the assessment, each *(search intent, document)* or *(recommendation intent, product)* pair was evaluated to ensure at least three human relevance judgments. Assessors had access to (i) the canonical formulation of the intent, (ii) the conversation up to the utterance from which the intent was derived, (iii) the textual description of the user profile (only for *recommendation intents*), and (iv) the document or product text. Based on this information, they assigned a relevance judgment on a 0-2 rating scale, defined as follows: **0 – Not Relevant:** the document or product is completely unrelated to the request for the considered user; **1 – Partially Relevant:** the document or product contains some information related to the query, including partial information or details about some particular facets of the topic, but does not provide a complete response; **2 - Highly Relevant:** The document or product is sufficient to provide a complete and meaningful response.

3. Limitations

IR experimental collections are typically created by exploiting IR systems to retrieve the documents later annotated by human assessors. Similarly, RS collections rely on historical data logs. In the CSR domain this is not possible as there exists no deployed system. This raises to a “chicken-and-egg” situation: the community lacks both CSR systems to extract the data from and data to develop CSR systems. Consequently, we were forced to build CoSRec treating and annotating search and recommendation intents separately. Since the conversations did not occur in a real-life scenario and were generated by an LLM, some utterances might feel unnatural to a human reader. Nevertheless, using CoSRec, the research community can develop CSR systems whose logs can be used as future collections.

4. Conclusions and Future Work

In this work, we introduced CoSRec, a novel dataset designed for the CSR context, which comprise 9.2k conversations encompassing pure search, pure recommendation, and mixed search-and-recommendation utterances, all generated using LLMs. A subset of 311 conversations has been human-annotated to evaluate their quality and to label user intents. Additionally, for 20 high-quality conversations, CoSRec provides relevance judgments for each labeled intent, personalized for recommendation scenarios. We believe that CoSRec will foster research in the area by providing a robust foundation for developing and evaluating CSR systems. To ensure reproducibility and encourage extensions, we make all code, scripts, prompts, and the dataset publicly available. Future work will focus on the generation and labeling of new conversations and the improvement of personalization, by including it in the generation process and extending it to the search intents. Furthermore, the current version of CoSRec will allow the development of actual integrated CSR systems that can be used to collect additional data, ground truth labels, and conversations.

Acknowledgments

This work has received support from CAMEO, n. 2022ZLL7MW PRIN 2022. Moreover, we acknowledge the support of “FAIR - Future Artificial Intelligence Research” - Spoke 1 ”Human-centered AI” (PE00000013), “Extreme Food Risk Analytics” (EFRA), GA n. 101093026, funded by the European Commission under the NextGeneration EU programme.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: grammar and spelling check, paraphrase, and reword. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] M. Alessio, S. Merlo, T. D. Noia, G. Faggioli, M. Ferrante, N. Ferro, C. I. Muntean, F. M. Nardini, F. Narducci, R. Perego, G. Santucci, N. Viterbo, Cosrec: A joint conversational search and recommendation dataset, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-17, 202, ACM, 2025. doi:10.1145/3726302.3730319.

[2] T. Di Noia, G. Faggioli, M. Ferrante, N. Ferro, F. Narducci, R. Perego, G. Santucci, CAMEO: fostering joint conversational search and recommendation, in: M. Atzori, P. Ciaccia, M. Ceci, F. Mandreoli, D. Malerba, M. Sanguinetti, A. Pellicani, F. Motta (Eds.), Proceedings of the 32nd Symposium of Advanced Database Systems, Villasimius, Italy, June 23rd to 26th, 2024, volume 3741 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 290–301. URL: https://ceur-ws.org/Vol-3741/paper33.pdf.

[3] Z. Si, Z. Sun, X. Zhang, J. Xu, X. Zang, Y. Song, K. Gai, J. Wen, When search meets recommendation: Learning disentangled search representation for recommendation, in: H. Chen, W. E. Duh, H. Huang, M. P. Kato, J. Mothe, B. Poblete (Eds.), Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023, ACM, 2023, pp. 1313–1323. URL: https://doi.org/10.1145/3539618.3591786. doi:10.1145/3539618.3591786.

[4] H. Zamani, W. B. Croft, Learning a joint search and recommendation model from user-item interactions, in: J. Caverlee, X. B. Hu, M. Lalmas, W. Wang (Eds.), WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020, ACM, 2020, pp. 717–725. URL: https://doi.org/10.1145/3336191.3371818. doi:10.1145/3336191.3371818.

[5] H. Zeng, S. Kallumadi, Z. Alibadi, R. F. Nogueira, H. Zamani, A personalized dense retrieval framework for unified information access, in: H. Chen, W. E. Duh, H. Huang, M. P. Kato, J. Mothe, B. Poblete (Eds.), Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023, ACM, 2023, pp. 121–130. URL: https://doi.org/10.1145/3539618.3591626. doi:10.1145/3539618.3591626.

[6] G. Penha, A. Vardasbi, E. Palumbo, M. D. Nadai, H. Bouchard, Bridging search and recommendation in generative retrieval: Does one task help the other?, in: T. D. Noia, P. Lops, T. Joachims, K. Verbert, P. Castells, Z. Dong, B. London (Eds.), Proceedings of the 18th ACM Conference on Recommender Systems, RecSys 2024, Bari, Italy, October 14-18, 2024, ACM, 2024, pp. 340–349. URL: https://doi.org/10.1145/3640457.3688123. doi:10.1145/3640457.3688123.

[7] S. Merlo, G. Faggioli, N. Ferro, A reproducibility study for joint information retrieval and recommendation in product search, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part IV, volume 15575 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 130–145. URL: https://doi.org/10.1007/978-3-031-88717-8_10. doi:10.1007/978-3-031-88717-8_10.

[8] P. Owoicho, J. Dalton, M. Aliannejadi, L. Azzopardi, J. R. Trippas, S. Vakulenko, TREC cast 2022: Going beyond user ask and system retrieve with initiative and response generation, in: I. Soboroff, A. Ellis (Eds.), Proceedings of the Thirty-First Text REtrieval Conference, TREC 2022, online, November 15-19, 2022, volume 500-338 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2022. URL: https://trec.nist.gov/pubs/trec31/papers/Overview_cast.pdf.

[9] J. Dalton, C. Xiong, J. Callan, TREC cast 2021: The conversational assistance track overview, in: I. Soboroff, A. Ellis (Eds.), Proceedings of the Thirtieth Text REtrieval Conference, TREC 2021, online, November 15-19, 2021, volume 500-335 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2021. URL: https://trec.nist.gov/pubs/trec30/papers/Overview-CAsT.pdf.

[10] J. Dalton, C. Xiong, J. Callan, Cast 2020: The conversational assistance track overview, in: E. M. Voorhees, A. Ellis (Eds.), Proceedings of the Twenty-Ninth Text REtrieval Conference,

TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020, volume 1266 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2020. URL: https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.C.pdf.

[11] J. Dalton, C. Xiong, J. Callan, TREC cast 2019: The conversational assistance track overview, CoRR abs/2003.13624 (2020). URL: https://arxiv.org/abs/2003.13624. arXiv:2003.13624.

[12] R. Li, S. E. Kahou, H. Schulz, V. Michalski, L. Charlin, C. Pal, Towards deep conversational recommendations, in: S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, 2018, pp. 9748–9758. URL: https://proceedings.neurips.cc/paper/2018/hash/800de15c79c8d840f4e78d3af937d4d4-Abstract.html.

[13] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng, MS MARCO: A human generated machine reading comprehension dataset, in: T. R. Besold, A. Bordes, A. S. d'Avila Garcez, G. Wayne (Eds.), Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016, volume 1773 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2016. URL: https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf.

[14] Y. Hou, J. Li, Z. He, A. Yan, X. Chen, J. J. McAuley, Bridging language and items for retrieval and recommendation, CoRR abs/2403.03952 (2024). URL: https://doi.org/10.48550/arXiv.2403.03952. doi:10.48550/ARXIV.2403.03952. arXiv:2403.03952.

[15] T. Liang, C. Jin, L. Wang, W. Fan, C. Xia, K. Chen, Y. Yin, LLM-REDIAL: A large-scale dataset for conversational recommender systems created from user behaviors with llms, in: L. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024, Association for Computational Linguistics, 2024, pp. 8926–8939. URL: https://doi.org/10.18653/v1/2024.findings-acl.529. doi:10.18653/V1/2024.FINDINGS-ACL.529.

[16] T. Formal, B. Piwowarski, S. Clinchant, SPLADE: sparse lexical and expansion model for first stage ranking, in: F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, T. Sakai (Eds.), SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, ACM, 2021, pp. 2288–2292. URL: https://doi.org/10.1145/3404835.3463098. doi:10.1145/3404835.3463098.

[17] S. Lin, J. Yang, J. Lin, Distilling dense representations for ranking using tightly-coupled teachers, CoRR abs/2010.11386 (2020). URL: https://arxiv.org/abs/2010.11386. arXiv:2010.11386.

[18] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, E. Grave, Unsupervised dense information retrieval with contrastive learning, Trans. Mach. Learn. Res. 2022 (2022). URL: https://openreview.net/forum?id=jKN1pXi7b0.