Engineering Conversational Search Systems: A Review of Applications, Architectures, and Functional Components

Phillip Schneider[1], Wessel Poelman[2], Michael Rovatsos[3], Florian Matthes[1]

16.08.2024, NLP4ConvAI @ ACL 2024

[1]Chair of Software Engineering for Business Information Systems
Department of Computer Science
School of Computation, Information and Technology (CIT)
Technical University of Munich

[2]Leuven AI Group of Multilingual NLP
Department of Computer Science
University of Leuven (KU Leuven)

[3]Artificial Intelligence and its Applications Institute (AIAI)
School of Informatics
University of Edinburgh

Outline

Introduction

- Conversational Information-Seeking
- Research Gap & Research Question

Method of Systematic Literature Review

Results & Discussion

- Definitions and Application Scenarios
- Architecture Framework
- Conversational Search Functions

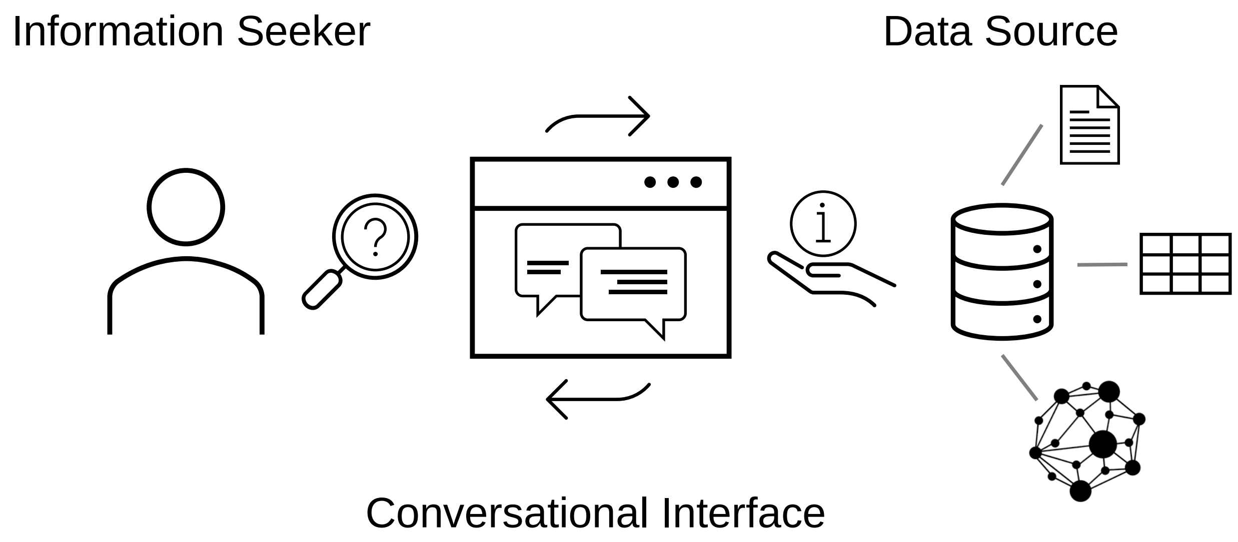
Conclusion & Future Outlook

Introduction: Conversational Information-Seeking

- **Conversational information-seeking** is an emerging **search paradigm** that frames information retrieval as interactive dialogues

- These conversational interfaces are often connected to very large data sources like **relational DBs, knowledge graphs,** or **document collections**

- Conversational information-seeking systems are usually distinguished into three categories (Zamani et al., 2023)
  - Conversational question-answering
  - Conversational recommendation
  - **Conversational search** (focus of this study)



Introduction: Research Gap & Research Question

- Growing body of research on conversational search is driven by the widespread adoption of **conversational interfaces** and the popularity surrounding **large language models** (LLMs) and retrieval augmented generation (**RAG**) systems

- However, there is a **lack of comprehensive reviews** and surveys in the literature:
  - Most studies have a narrow focus on specific technical functions or application domains
  - Apparent **gap between theoretical frameworks** and **actual implementations**

- We provide a **system-centric review across the development process**, ranging from conceptualizing system functionalities to implementing architectural components

Main Research Question

What are the suitable application scenarios, established system architectures, and core functional components for developing conversational search systems?

Method of Systematic Literature Review

- We conducted a **systematic literature review** based on the guidelines from Kitchenham (Kitchenham et al.,2004)
- To obtain relevant publications, we applied our search string to query **six academic databases** from **2012-2022**, yielding a final set of 51 papers that met our selection criteria
- In addition, we used **forward snowballing** to identify recent papers from **2023 and 2024** that focus on augmenting conversational search systems with **LLMs**
- Our paper also lists **16 datasets** that are commonly used in the identified papers

Search String

“conversational search” OR

“information-seeking dialogue” OR

“conversational information retrieval” OR

“conversational information-seeking” OR

“information-seeking conversation"

Academic Databases

- **ACL Anthology**
- **ACM Digital Library**
- **IEEE Xplore**
- **ScienceDirect**
- **Scopus**
- **Web of Science**

Results & Discussion: Definitions and Application Scenarios

- While the overarching goal of conversational search remains consistent, scholars define conversational search systems from three distinct perspectives: **dialogue-**, **system-**, and **task-oriented definitions**

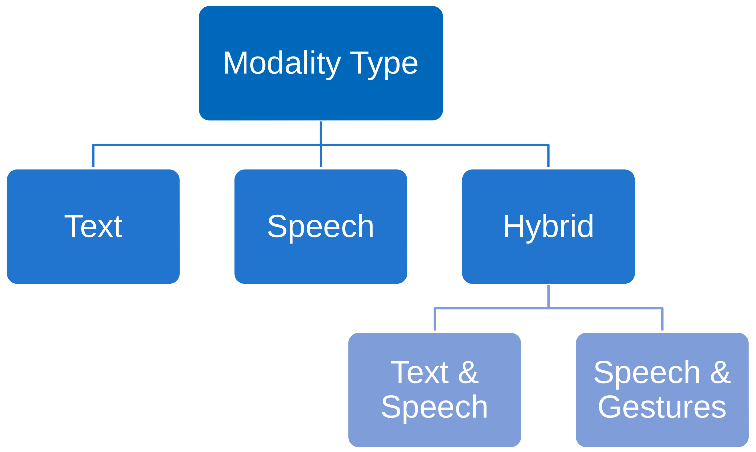
- Despite focusing on different aspects, these definitions highlight **four key system properties** that distinguish conversational search systems from classic search systems
(similar to Radlinski and Craswell (2017))

System Goal	Maximize a user's information gain during a dialogue by providing search results with maximum utility.			
System Definitions	Dialogue-oriented	System-oriented	Task-oriented	
System Properties	Mixed-initiative interaction	Mutual understanding	Context & memory	Continuous refinement

Results & Discussion: Definitions and Application Scenarios

- The suitability of conversational search depends on the **search task** and **search modality**
- Conversational search excels in **complex search**, ambiguous scenarios with **iterative clarifications** and **feedback loops**, rather than straightforward known-item searches (Radlinski and Craswell, 2017)
- Conversational search systems can support **text-based**, **speech-based**, or **hybrid** interactions, yet most systems are unimodal and text-based, but multimodal systems are on the rise (Liao et al., 2021)

Observed Modalities



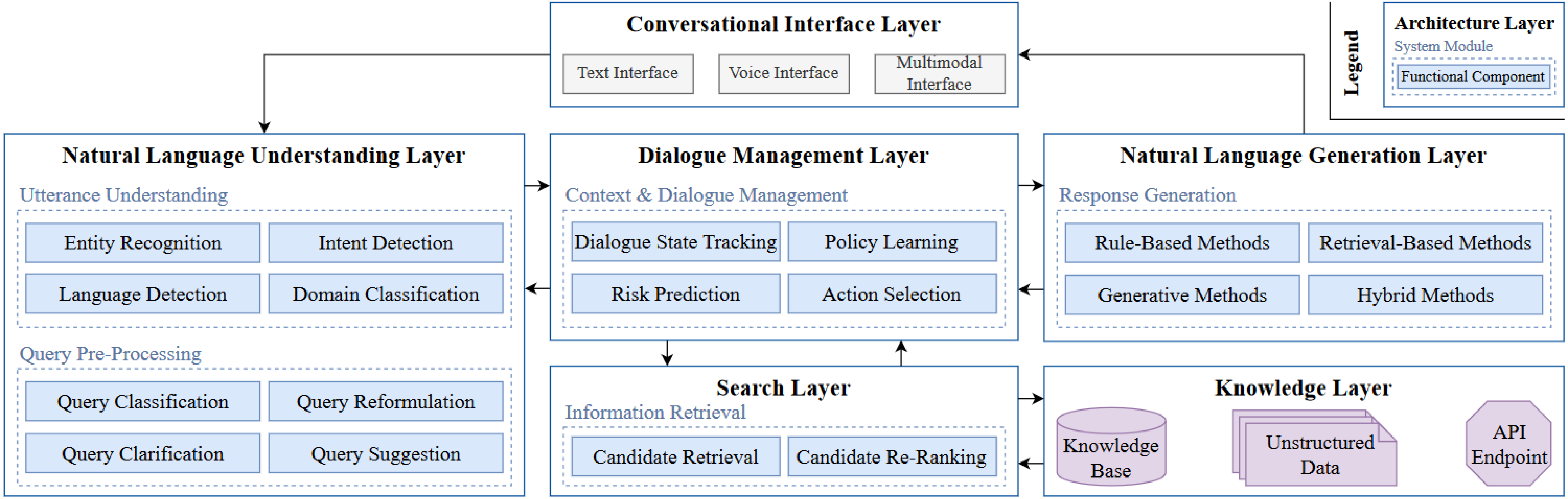
Most Popular Application Domains

Business	Health	Law	Tourism
Gastronomy	Media	Public services	Aerospace

- Domain-specific systems help users to initiate a search **without prior domain knowledge** and assist when certain **modalities are restricted**

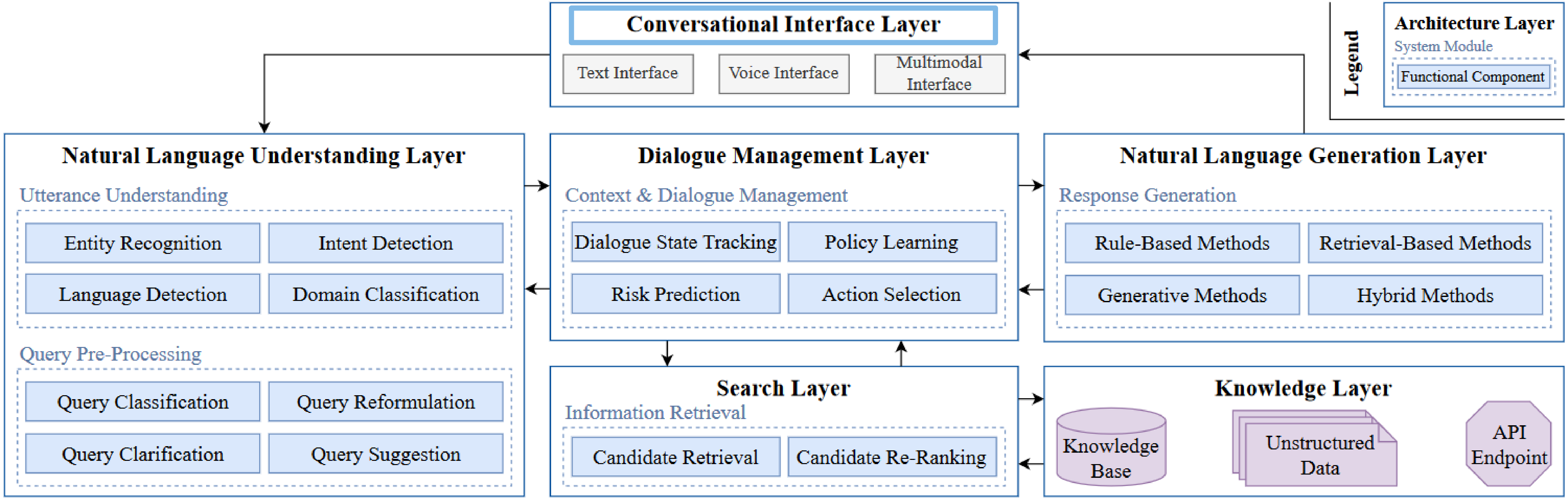
Results & Discussion: Architecture Framework

- Based on over 20 architectures proposed in the literature, we devised a **layered architectural framework** for conversational search systems
- A combination of the set of **modules** and **functional components** implements the stated system properties



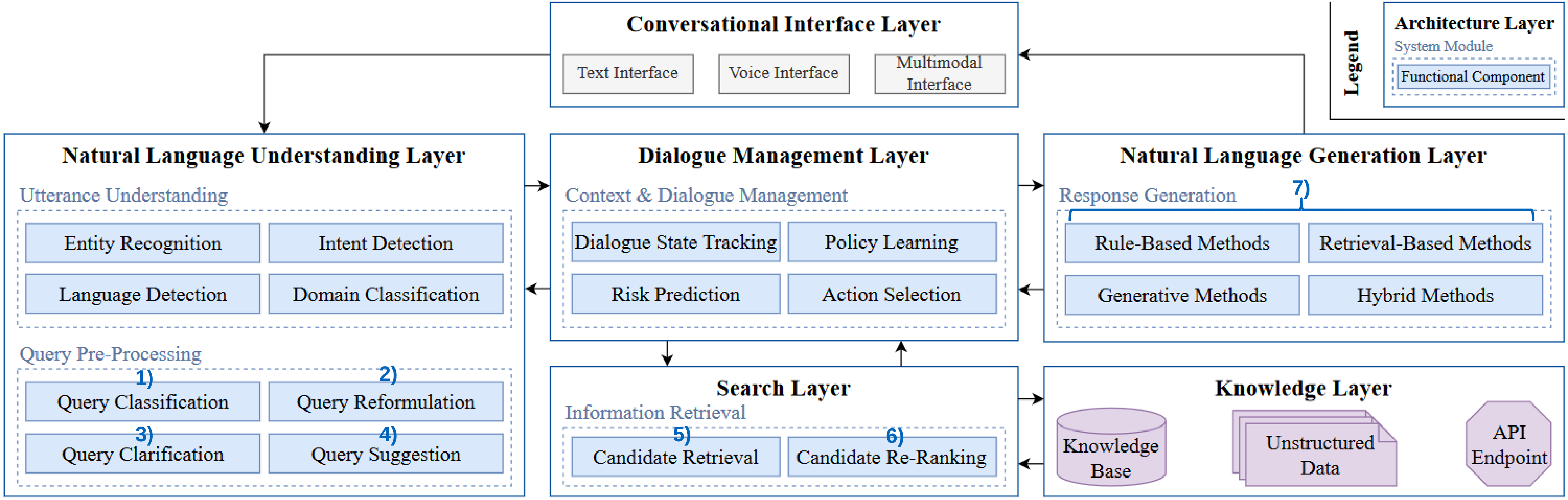
Results & Discussion: Architecture Framework

- Based on over 20 architectures proposed in the literature, we devised a **layered architectural framework** for conversational search systems
- A combination of the set of **modules** and **functional components** implements the stated system properties



Results & Discussion: Architecture Framework

- Based on over 20 architectures proposed in the literature, we devised a **layered architectural framework** for conversational search systems
- A combination of the set of **modules** and **functional components** implements the stated system properties



1) – 7) Core Conversational Search Functions

Results & Discussion: Conversational Search Functions

1) Query Classification

Function Description

- Classify a given user query to inform subsequent functions, addressing issues regarding ambiguity, domain-specificity, or other contextual aspects.

Approaches

- Classifying **search domains** and domain-specific information needs can aid in choosing the correct data source and answer format (Frummet et al., 2019; Hamzei et al., 2020)
- Classifying whether a **previous question is relevant** to the current question can provide additional context (Aliannejadi et al., 2020)
- Classifying **question types** by using question words and keywords can help to give more appropriate answers (Kia et al., 2020)

History
Who formed **Saosin**?
When was the **band** founded?
What was their **first** album?
Query
When was the album released?
Relevant Answer Passage
The original lineup for Saosin, consisting of Burchell, Shekoski, Kennedy and Green, was formed in the summer of 2003. On June 17, the band released their first commercial production, the EP Translating the Name.

Example excerpt from conversational search dialogue (Voskarides et al., 2020)

Results & Discussion: Conversational Search Functions

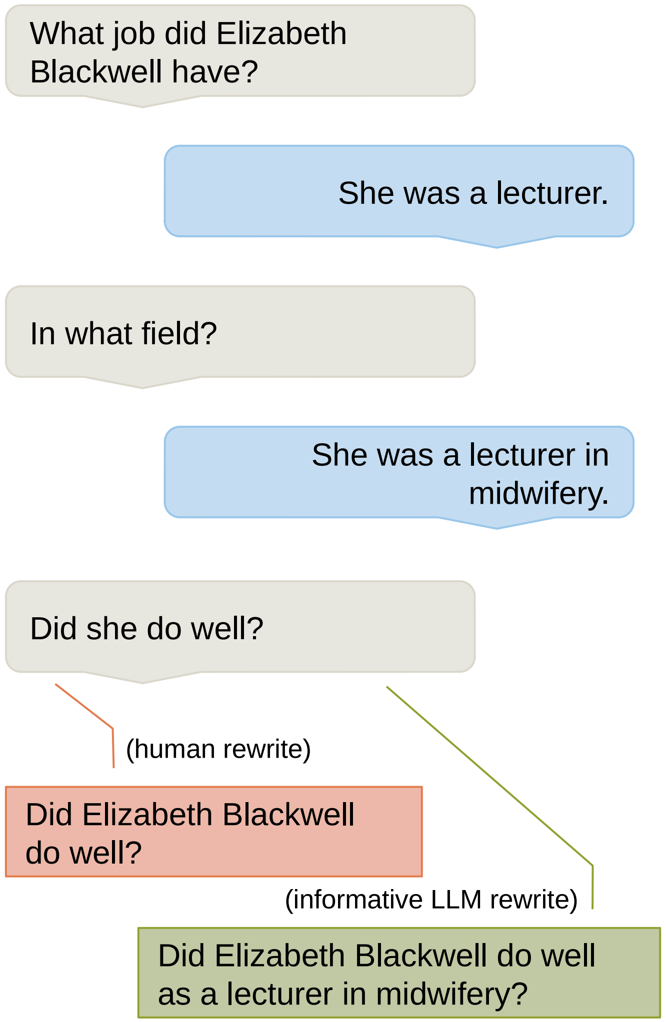
2) Query Reformulation

Function Description

- Rewrite ambiguous user queries into a clear, explicit form with more contextual information for better downstream retrieval performance.

Approaches

- Classifying terms to be included in the rewritten query as well as sequence-to-sequence rewriting approaches for **co-reference resolution** and **query expansion** (Mele et al., 2021; Zhang et al., 2021)
- LLMs have been used to iteratively rewrite the query in real-time until they align with the user‘s intent and also to incorporate **additional information** to decontextualize the query (Chen et al., 2023; Ye et al., 2023)



Example of informative query rewriting (Ye et al., 2023)

Results & Discussion: Conversational Search Functions

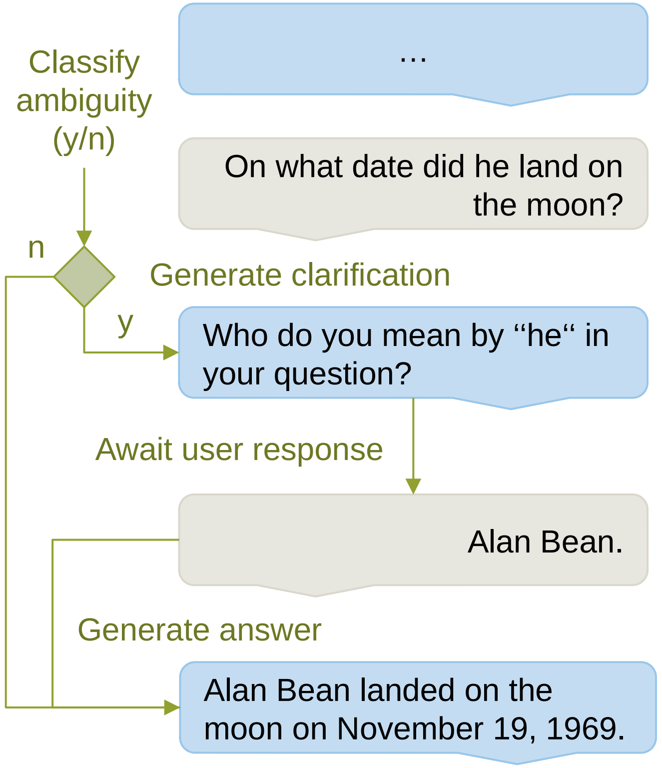
3) Query Clarification

Function Description

- Take the initiative to proactively ask the user for clarification if the system cannot interpret or resolve a given query.

Approaches

- Different approaches have been investigated, including template filling, sequence editing, sequence-to-sequence, or hybrid approaches (Zamani et al., 2020)
- However, these approaches have to carefully **balance information gain** and **user patience** (Bi et al., 2021)
- Prompting LLMs can be used to detect first whether a given question is **ambiguous** and then **generate** an **appropriate clarification** question to ask the user (Kuhn et al., 2023)



CLAM framework for selective clarification with LLMs (Kuhn et al., 2023)

Results & Discussion: Conversational Search Functions

4) Query Suggestion

Function Description

- Take the initiative to proactively suggest relevant queries or (partial) answers during the conversational interaction with the user.

Approaches

- Most approaches employ transformer models and use the dialogue history and input query to **generate** and **rank suggestions**, aiming to maximize the probability of a user picking one of the suggestions (Dehghani et al., 2017; Mustar et al., 2022)
- Methods can be implemented as **auto-complete** functions or **listing** suggestions
- Additionally, LLMs can be employed to **generate multiple suggestions** that users can iteratively accept, edit, or expand (Anand et al., 2023)
- However, as with clarifications, **suggestions** can tend to have **diminishing returns** on the information gain (Aliannejadi et al., 2021)

Gold Label: Misses Intent
Query: used washer and dry
Question Suggestion: Can I store a washer and dryer in the garage?
Gold Label: Prequel
Query: verizon yahoo purchase
Question Suggestion: Who bought out Yahoo?
Gold Label: Too specific
Query: medicaid expansion
Question Suggestion: Did Florida accept Medi-caid expansion?
Gold Label: Useful
Query: best hair clippers
Question Suggestion: What clippers do barbers use?

Examples of query-question suggestion pairs and their usefulness labels (Rosset et al., 2020)

Results & Discussion: Conversational Search Functions

5) Candidate Retrieval

Function Description

- Fetch the most relevant data items for a given user query from a structured DB, unstructured text collections, or a semi-structured data source.

Approaches

- Two main approaches for retrieving information from unstructured text collections
  - **Sparse retrieval** uses methods like BM25, relying on sparse vectors to encode term occurrences in queries and documents (Robertson and Zaragoza, 2009)
  - **Dense retrieval** addresses the limitations of sparse retrieval using transformer-based encoder models and dense vectors (Ferreira et al., 2022)
- LLMs can be used for **generating synthetic training data** for dense retrieval models (Huang et al., 2023)
- LLMs have shown to be capable in the task of **semantic parsing**, producing structured DB queries for a given question and dialogue (Schneider et al., 2024a)

System Prompt

- **SYSTEM**:
  Generate a SPARQL query that answers the given 'Input question:'. […]

Few-Shot Example

- **USER**:
  Conversation history:
  USER: Which administrative territory is the native country of Cirilo Villaverde ?
  SYSTEM: {'Q241': 'Cuba'}

  Input question: Which is the national anthem of that administrative territory ?

  Entities: {'Q241': 'Cuba'}
  Relations: {'P85': 'anthem'}
  Types: {'Q484692': 'hymn'}

  ASSISTANT:
  SPARQL query: SELECT ?x WHERE { wd:Q241 wdt:P85 ?x . ?x wdt:P31 wd:Q484692 . }

Input Prompt

- **USER**:
  Conversation history:
  <conversation_history>
  Input question: *<utterance>*
  Entities: *<entities>*
  Relations: *<relations>*
  Types: *<types>*

Example of few-shot prompting of LLM for semantic parsing in conversational QA (Schneider et al., 2024a)

Results & Discussion: Conversational Search Functions

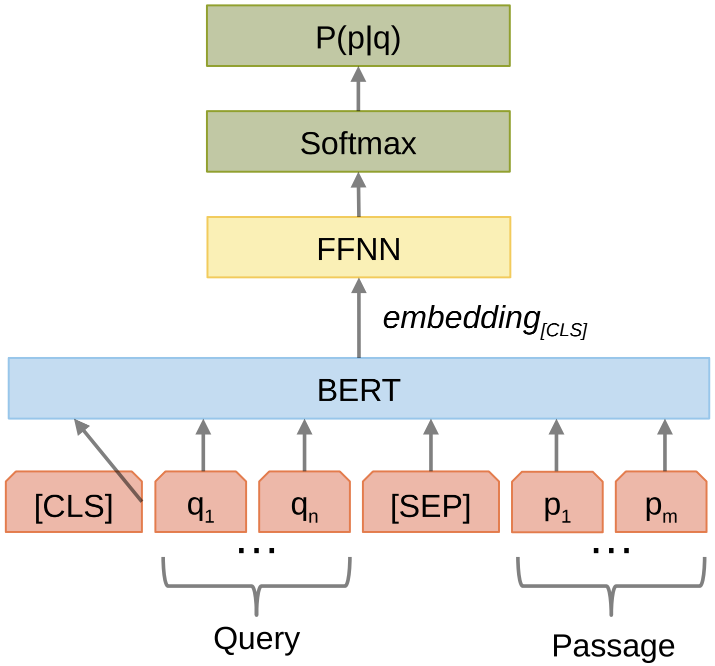
6) Candidate Re-Ranking

Function Description

- Rank retrieved candidates in order of informativeness and relevancy with regard to the given user query.

Approaches

- Most methods are training models to **score and reorder** candidates, incorporating various embeddings and dialogue history by training on query-item pairs or using distance measures (Ferreira et al., 2022)
- More elaborate approaches can use **multiview re-ranking** where multiple embeddings of the input query can be used for an aggregated ranking (Kumar and Callan, 2020)
- Other systems use multiple LLM-based agents where a dialogue manager and a re-ranker reorder candidates and also **generate explanations** (Friedman et al., 2023)



Example of simple BERT-based re-ranker architecture (Ferreira et al., 2022)

Results & Discussion: Conversational Search Functions

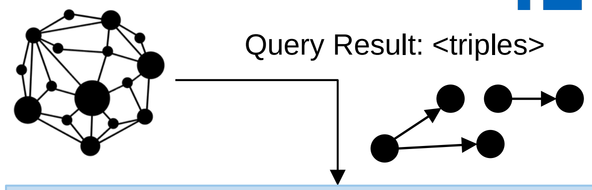
7) Knowledge-Based Response Generation

Function Description

- Generate a relevant natural language response based on retrieved information from the knowledge layer.

Approaches

- Most approaches depend on three categories: **information type**, **generation method**, and **information source** (Zamani et al., 2023)
- Common methods are template filling, sequence-to-sequence methods, and **RAG**-based methods (Zhang et al., 2018; Ferreira et al., 2022; Lewis et al., 2020; Shuster et al., 2021)
- LLMs are especially capable in RAG, even when using **structured data** as input like **semantic triples** from knowledge graphs (Schneider et al., 2024b)



(Fine-Tuned) LLM

System Prompt

- **SYSTEM**:
Generate a concise text for the given set of triples. Ensure that the generated output only includes […]

Few-Shot Example

- **USER**:
Input triples:
[{'object': 'Albert_E._Austin', 'property': 'successor', 'subject': 'Alfred_N._Phillips'}, {'object': 'Connecticut', 'property': 'birthPlace', 'subject': 'Alfred_N._Phillips'}, {'object': 'United_States_House_of_Representatives', 'proper ty': 'office', 'subject': 'Alfred_N._Phillips'}]

ASSISTANT:
Output text:
Albert E. Austin succeeded Alfred N. Phillips who was born in Connecticut and worked at the United States House of Representatives.

Input Prompt

- **USER**:
Input triples: <triples>

Example of few-shot prompting of LLM for semantic triple verbalization (Schneider et al., 2024a)

Conclusion & Future Outlook

- **Rise in Popularity and Diversification:** Conversational search systems are becoming more and more popular, with significant diversification in interaction modalities and application domains

- **Conversational Search Architecture Framework:** We consolidate a generalized architecture framework based on validated systems from the literature

- **Rapid Adoption of LLMs:** Researchers increasingly incorporate LLMs, particularly in replacing classic NLU pipelines, often using prompting instead of fine-tuning; however, challenges like model size, hallucinations, and a lack of transparency as well as controllability persist

- **Augmentation, Not Replacement:** LLMs are unlikely to replace modular conversational search systems as a single end-to-end solution; instead, they augment the functions of the proposed modular framework

- **Future Outlook:** The trend is shifting towards function-specific, smaller LLMs that complement existing system components, rather than developing a monolithic model to handle all conversational search functions

Paper Link
https://arxiv.org/abs/2407.00997

References

Aliannejadi, M., Chakraborty, M., Ríssola, E. A., & Crestani, F. (2020). Harnessing Evolution of Multi-Turn Conversations for Effective Answer Retrieval. *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, 33–42. https://doi.org/10.1145/3343413.3377968

Anand, A., V, V., Anand, A., & Setty, V. (2023). Query Understanding in the Age of Large Language Models. *Gen-IR@SIGIR 2023: The First Workshop on Generative Information Retrieval*. https://doi.org/10.48550/arXiv.2306.16004

Bi, K., Ai, Q., & Croft, W. B. (2021). Asking Clarifying Questions Based on Negative Feedback in Conversational Search. *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, 157–166. https://doi.org/10.1145/3471158.3472232

Chen, Z., Jiang, Z., Yang, F., Cho, E., Fan, X., Huang, X., Lu, Y., & Galstyan, A. (2023). Graph Meets LLM: A Novel Approach to Collaborative Filtering for Robust Conversational Understanding. In M. Wang & I. Zitouni (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track* (pp. 811–819). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-industry.75

Dehghani, M., Rothe, S., Alfonseca, E., & Fleury, P. (2017). Learning to Attend, Copy, and Generate for Session-Based Query Suggestion. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1747–1756. https://doi.org/10.1145/3132847.3133010

Ferreira, R., Leite, M., Semedo, D., & Magalhaes, J. (2022). Open-Domain Conversational Search Assistants: The Transformer Is All You Need. *Information Retrieval*, *25*(2), 123–148. https://doi.org/10.1007/s10791-022-09403-0

Friedman, L., Ahuja, S., Allen, D., Tan, T., Sidahmed, H., Long, C., Xie, J., Schubiner, G., Patel, A., Lara, H., & others. (2023). Leveraging Large Language Models in Conversational Recommender Systems. *ArXiv Preprint ArXiv:2305.07961v2*. https://doi.org/10.48550/arXiv.2305.07961

Frummet, A., Elsweiler, D., & Ludwig, B. (2019, November). Detecting Domain-Specific Information Needs in Conversational Search Dialogues. *Natural Language for Artificial Intelligence*. https://ceur-ws.org/Vol-2521/paper-02.pdf

Hamzei, E., Li, H., Vasardani, M., Baldwin, T., Winter, S., & Tomko, M. (2020). Place Questions and Human-Generated Answers: A Data Analysis Approach. In P. Kyriakidis, D. Hadjimitsis, D. Skarlatos, & A. Mansourian (Eds.), *Geospatial Technologies for Local and Regional Development* (pp. 3–19). Springer International Publishing. https://doi.org/10.1007/978-3-030-14745-7_1

Huang, C. W., Hsu, C. Y., Hsu, T. Y., Li, C. A., & Chen, Y. N. (2023). CONVERSER: Few-shot conversational dense retrieval with synthetic data generation. arXiv preprint arXiv:2309.06748.

Kia, O. M., Neshati, M., & Alamdari, M. S. (2020). Open-Domain Question Classification and Completion in Conversational Information Search. *2020 11th International Conference on Information and Knowledge Technology (IKT)*, 98–101. https://doi.org/10.1109/IKT51791.2020.9345613

Kitchenham, B. A., Dyba, T., & Jorgensen, M. (2004). Evidence-Based Software Engineering. *Proceedings of the 26th International Conference on Software Engineering*, 273–281. https://doi.org/10.1109/ICSE.2004.1317449

Kuhn, L., Gal, Y., & Farquhar, S. (2023). CLAM: Selective Clarification for Ambiguous Questions with Generative Language Models. *ICML 2023 Workshop on Deployment Challenges for Generative AI*.

Kumar, V., & Callan, J. (2020). Making Information Seeking Easier: An Improved Pipeline for Conversational Search. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3971–3980. https://doi.org/10.18653/v1/2020.findings-emnlp.354

References

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., & others. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Liao, L., Long, L. H., Zhang, Z., Huang, M., & Chua, T.-S. (2021). MMConv: An Environment for Multimodal Conversational Search across Multiple Domains. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 675–684. https://doi.org/10.1145/3404835.3462970

Mele, I., Muntean, C. I., Nardini, F. M., Perego, R., Tonellotto, N., & Frieder, O. (2021). Adaptive Utterance Rewriting for Conversational Search. *Information Processing and Management: An International Journal*, *58*(6). https://doi.org/10.1016/j.ipm.2021.102682

Mustar, A., Lamprier, S., & Piwowarski, B. (2022). On the Study of Transformers for Query Suggestion. *ACM Transactions on Information Systems*, *40*(1), 1–27. https://doi.org/10.1145/3470562

Radlinski, F., & Craswell, N. (2017). A Theoretical Framework for Conversational Search. *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, 117–126. https://doi.org/10.1145/3020165.3020183

Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*, *3*(4), 333–389. https://doi.org/10.1561/1500000019

Schneider, P., Klettner, M., Jokinen, K., Simperl, E., & Matthes, F. (2024a). Evaluating Large Language Models in Semantic Parsing for Conversational Question Answering over Knowledge Graphs. *International Conference on Agents and Artificial Intelligence*, 807–814. https://doi.org/10.5220/0012394300003636

Schneider, P., Klettner, M., Simperl, E., & Matthes, F. (2024b). A Comparative Analysis of Conversational Large Language Models in Knowledge-Based Text Generation. In Y. Graham & M. Purver (Eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 358–367). Association for Computational Linguistics. https://aclanthology.org/2024.eacl-short.31

Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval Augmentation Reduces Hallucination in Conversation. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3784–3803. https://doi.org/10.18653/v1/2021.findings-emnlp.320

Voskarides, N., Li, D., Ren, P., Kanoulas, E., & de Rijke, M. (2020, July). Query resolution for conversational search with limited supervision. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval* (pp. 921-930).

Ye, F., Fang, M., Li, S., & Yilmaz, E. (2023). Enhancing Conversational Search: Large Language Model-Aided Informative Query Rewriting. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 5985–6006). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.398

Zamani, H., & Craswell, N. (2020). Macaw: An Extensible Conversational Information Seeking Platform. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2193–2196. https://doi.org/10.1145/3397271.3401415

Zamani, H., Trippas, J. R., Dalton, J., & Radlinski, F. (2023). Conversational Information Seeking. *ArXiv Preprint ArXiv:2201.08808v2*.

Zhang, E., Lin, S.-C., Yang, J.-H., Pradeep, R., Nogueira, R., & Lin, J. (2021). Chatty Goose: A Python Framework for Conversational Search. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2521–2525. https://doi.org/10.1145/3404835.3462782

Zhang, Y., Chen, X., Ai, Q., Yang, L., & Croft, W. B. (2018). Towards Conversational Search and Recommendation: System Ask, User Respond. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 177–186. https://doi.org/10.1145/3269206.3271776

Phillip Schneider

Technical University of Munich (TUM)
TUM School of CIT
Department of Computer Science (CS)
Chair of Software Engineering for Business
Information Systems (sebis)

Boltzmannstraße 3
85748 Garching bei München