# Answering Topical Information Needs Using Neural Entity-Oriented Information Retrieval and Extraction

Shubham Chatterjee
University of New Hampshire
USA
shubham.chatterjee@unh.edu

**Abstract**

In the modern world, search engines are an integral part of human lives. The field of Information Retrieval (IR) is concerned with finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need (query) from within large collections (usually stored on computers). The search engine then displays a ranked list of results relevant to our query. Traditional document retrieval algorithms match a query to a document using the overlap of words in both. However, the last decade has seen the focus shifting to leveraging the rich semantic information available in the form of *entities*.

Entities are uniquely identifiable objects or things such as places, events, diseases, etc. that exist in the real or fictional world. Entity-oriented search systems leverage the semantic information associated with entities (e.g., names, types, etc.) to better match documents to queries. Web search engines would provide better search results if they understand the meaning of a query.

This dissertation advances the state-of-the-art in IR by developing novel algorithms that understand text (query, document, question, sentence, etc.) at the semantic level. To this end, this dissertation aims to understand the fine-grained meaning of entities from the context in which the entities have been mentioned, for example, "oysters" in the context of food versus ecosystems. Further, this dissertation aims to automatically learn (vector) representations of entities that incorporate this fine-grained knowledge and knowledge about the query. This dissertation refines the automatic understanding of text passages using deep learning, a modern artificial intelligence paradigm.

This dissertation utilizes the semantic information extracted from entities to retrieve materials (text and entities) relevant to a query. The interplay between text and entities in the text is studied by addressing three related prediction problems: (1) Identify entities that are relevant for the query, (2) Understand an entity's meaning in the context of the query, and (3) Identify text passages that elaborate the connection between the query and an entity.

The research presented in this dissertation may be integrated into a larger system designed for answering complex topical queries such as *dark chocolate health benefits* which require the search engine to automatically understand the connections between the query and the relevant material, thus transforming the search engine into an answering engine.

**Awarded by:** University of New Hampshire, Durham, USA **on** 1 September 2022.
**Supervised by:** Laura Dietz.
**Available at:** https://scholars.unh.edu/dissertation/2714/.

## Selected Publications


Shubham Chatterjee and Laura Dietz. Why Does This Entity Matter? Support Passage Retrieval for Entity Retrieval. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, page 221–224, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368810. doi: 10.1145/3341981.3344243. URL https://doi.org/10.1145/3341981.3344243.

Shubham Chatterjee and Laura Dietz. Entity Retrieval Using Fine-Grained Entity Aspects. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 1662–1666, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3463035. URL https://doi.org/10.1145/3404835.3463035.

Shubham Chatterjee and Laura Dietz. BERT-ER: Query-Specific BERT Entity Representations for Entity Ranking. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 1466–1477, New York, NY, USA, 2022a. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3531944. URL https://doi.org/10.1145/3477495.3531944.

Shubham Chatterjee and Laura Dietz. Predicting Guiding Entities for Entity Aspect Linking. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, CIKM '22, New York, NY, USA, 2022b. Association for Computing Machinery. ISBN 978145039236. doi: 10.1145/3511808.3557671. URL https://doi.org/10.1145/3511808.3557671.

Laura Dietz, Shubham Chatterjee, Connor Lennox, Sumanta Kashyapi, Pooja Oza, and Ben Gamari. Wikimarks: Harvesting Relevance Benchmarks from Wikipedia. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3003–3012, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3531731. URL https://doi.org/10.1145/3477495.3531731.