Improving Web Search by Identifying Fine-Grained Entity Information in Text

Shubham Chatterjee

Advisor: Dr. Laura Dietz

Research proposal submitted in the partial fulfillment of the requirements for

Dissertation Year Fellowship of the University of New Hampshire

College of Engineering and Physical Sciences

Department Of Computer Science

ABSTRACT

Web search engines could provide better (and less redundant) search results, if they would be able to understand the meaning of relevant information better. One way of doing this is by having a deeper understanding of entities (people, organizations, events, places, etc.) in text. Through this project, my goal is to develop algorithms which can infer different events, topics, roles, and more generally, different aspects of an entity from the context in which it is mentioned. My approach refines the automatic understating of text passages using deep learning, a modern artificial intelligence paradigm. My approach identifies relevant words and relevant entities simultaneously and provides more fine-grained information about entities in some text to help machines understand the meaning of the text better.

1. Introduction. Automatic algorithms for text understanding are essential for many artificial intelligence (AI) tasks. For example, in web search, search engines need to understand the text from the web pages. Such web pages are retrieved based on the overalap of words between the user's information need (henceforth called *query*) and the web page [6, 14, 13]. However, during the last decade, it has become popular to also use the rich semantic information available in the form of *entities* in text.

Entities are uniquely identifiable objects or things, such as people, events, diseases, etc. Some example entities are shown in bold italics in Figure 1. Entities are important for many search tasks such as answering factoid questions like *What is the capital of India?* It has been estimated that over 40% of Web search queries target entities [11, 7].

With the increasing importance of entities in web search, tools [5, 9] which identify mentions of entities in text have been developed. Given some text such as the one on the left in Figure 1, such tools can identify that the mention "oyster" refers to the animal and not to the place [1] in Virginia, United States. However, it may be more beneficial for a user who is researching the the role of oysters in ecosystems, to know that the entity "oyster" has been mentioned in the context of its role as an ecosystem engineer, and not its cultivation. Hence, there is a need for better tools which are able to understand text at a deeper level. In this regard, for my dissertation year, I propose to work on the following problem:

Entity Aspect Linking (EAL). Given a mention of an entity in a context such a tweet, sentence or paragraph, and a set of predefined aspects with their associated content, link the mention with an aspect that best captures the addressed topic.

The outcome of this task is refined knowledge about an entity, which provides richer information about the entity and enables the user to choose the granularity appropriate for the task at hand. I will evaluate the efficacy of my research using an established benchmark from Ramsdell et al. [12].

Challenges in Entity Aspect Linking for a Machine. Identifying which aspects of an entity are referenced in a context is easy for humans. When given a relevant text passage, humans

[1] https://en.wikipedia.org/wiki/Oyster,_Virginia

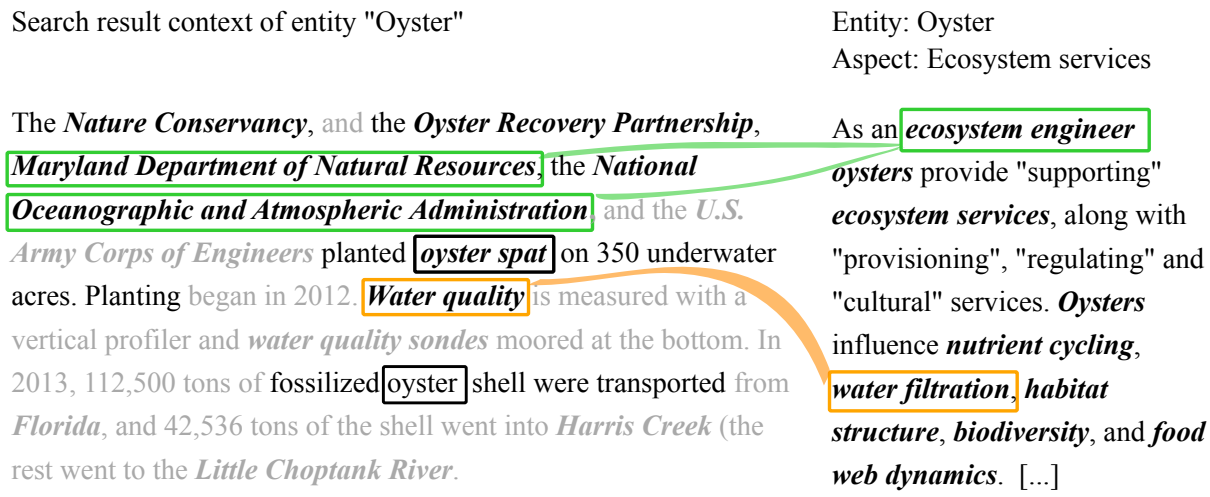

Figure 1: Depiction of our automated approach for identifying the correct aspect of the entity "oyster". Left: context from search results. Right: Correct aspect "Ecosystem services" of the entity "Oyster". The example text, entities, and aspects are taken from Wikipedia. Mentioned entities, i.e., entity links, are marked in bold italics. In objective O1, we address the issue that not all words are relevant for the decision - non-relevant words are depicted in grey. As described in objective O2, it is rare that identical entities are mentioned in both context and aspect content, hence we need to identify which entities are related in this context, such as entities related to ecosystems (green frame) and regarding water quality (orange frame). In objective O3, we elaborate how integrating the prediction of relevant words and entities is helpful for most accurate predictions of entity aspect links.

(1) focus on most relevant words in the context, (2) can infer which entities are related to other entities in the context, and (3) know how the words and entities connect to each other. My goal is to develop machine learning algorithms that would teach a machine to accomplish the above tasks.

2. Proposed Research Activities.

Concrete Objectives. During my dissertation year, I will be focusing on the following research objectives:

O1 Identify which words/segments from the aspect's content is relevant for the entity we are trying to aspect-link, as typically, only some words/segments are relevant and useful.

O2 Identify pertinent connections between entities from the aspect's content and the entity's context which might be useful for aspect linking decisions.

O3 Integrate words and entities, considering that many words from both, the entity's context and the aspect's content, are non-relevant, and that the pertinence of entity connections

are context dependent.

Previous Work and Pilot Studies. My proposed research is based on lessons learned from related work and my own pilot studies, as described below.

O1 Nanni et al.[10] find that using all words from the entity's context leads to poor results, and alleviate this issue by considering only the sentence mentioning the entity. However, a sentence may not always provide us with the entire context to help us make the aspect linking decision. For example, in Figure 1, we need to consider the whole passage which mentions the entity "oyster", and not just the sentence. Since the majority of words in the larger context are not relevant, we must be able to identify only the relevant words from a larger context.

O2 Findings from my previous work [1] show that it is important to consider the user's query to find relevant fine-grained information about an entity. Previous work [10] has based the aspect linking decisions on whether a direct relationship exists between an aspect-entity and a context-entity. However, as shown in Figure 1, otherwise unrelated entities such as "Ecosystem Engineer" and "Oceanographic and Atmospheric Administration" are related in the given context. Hence, in my work, I would base the aspect lining decisions on whether two entities are related in context.

O3 Previous works which leverage entities for retrieving text [2, 15, 8, 16, 4, 17] have found that combining indicators of relevance obtained using words and entities leads to better performance for distinguishing relevant from non-relevant text. In this light, I aim to integrate the information from relevant words and entities (from O1 and O2 above).

Proposed Approach and Implementation.

O1 My aim is to find words from the entity-context which are important for the aspect linking decision. My algorithmic approach would be to use a novel paradigm in deep learning called *attention*. Attention is based on the intuition that we "attend to" a certain part when processing a large amount of information. Using attention mechanisms in deep

neural networks, we can teach machines to select words that are most beneficial when included in the aspect linking decision.

O2 My aim is to consider the context, to find whether two entities are related. My algorithmic approach would be to find a representation of these entities which includes the context. This can be achieved by creating *embeddings* of these entities which include the context. An *embedding* is a numerical representation of words/entities that allows words/entities with similar meaning to have a similar representation. To create such entity embeddings, I would use a state-of-the-art embedding generator called BERT [3], albeit some modifications (which is the purpose of this research).

O3 My aim is to consider how the words and the entities from text interact with each other. My algorithmic approach would be to use a deep learning method called Siamese Neural Network. Siamese networks contain two deep learning networks which are exactly the same, and which work in parallel. I plan to extend this network to have four, instead of two parallel networks. The first network would consider all aspect entities, the second would consider all aspect words, the third would consider all context entities, and the fourth would consider all context words. This would inform me how the words/entities from the aspect interact with the words/entities from the entity's context.

3. Conclusion. In this proposal, I describe my research on entity aspect linking and identify three concrete objectives: (1) identifying words from the context and aspects which are relevant for aspect linking, (2) finding relatedness between two entities considering the context provided, and (3) basing the aspect linking decisions on the interactions of the relevant words from (1), and entities from (2).

Although we study the problem in the context of web search, applications such as question-answering systems and recommender systems which aim to understand the subtleties in human language would also benefit from this research. Entity aspect links would not only allow machines to achieve a deeper understanding of human language, but also allow users to understand fine-grained connections between entities through their aspects.

References

[1] Chatterjee, S., and Dietz, L. 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* (New York, NY, USA, 2019), ICTIR '19, Association for Computing Machinery, p. 221–224.

[2] Dalton, J., Dietz, L., and Allan, J. Entity query feature expansion using knowledge base links. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2014), SIGIR '14, Association for Computing Machinery, p. 365–374.

[3] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, Minnesota, June 2019), Association for Computational Linguistics, pp. 4171–4186.

[4] Ensan, F., and Bagheri, E. Document retrieval model through semantic linking. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining* (New York, NY, USA, 2017), WSDM '17, Association for Computing Machinery, p. 181–190.

[5] Ferragina, P., and Scaiella, U. Tagme: On-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (New York, NY, USA, 2010), CIKM '10, Association for Computing Machinery, p. 1625–1628.

[6] Jones, K. S., Walker, S., and Robertson, S. E. A probabilistic model of information retrieval: Development and comparative experiments part 2. *Information Processing and Management 36*, 6 (Nov. 2000), 809–840.

[7] Lin, T., Pantel, P., Gamon, M., Kannan, A., and Fuxman, A. Active objects: Actions for entity-centric search. In *Proceedings of the 21st International Conference on World Wide*

Web (New York, NY, USA, 2012), WWW '12, Association for Computing Machinery, p. 589–598.

[8] Liu, X., and Fang, H. Latent entity space: a novel retrieval approach for entity-bearing queries. *Information Retrieval Journal 18*, 6 (2015), 473–503.

[9] Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems* (New York, NY, USA, 2011), I-Semantics '11, Association for Computing Machinery, p. 1–8.

[10] Nanni, F., Ponzetto, S. P., and Dietz, L. Entity-aspect linking: Providing fine-grained semantics of entities in context. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries* (New York, NY, USA, 2018), JCDL '18, Association for Computing Machinery, p. 49–58.

[11] Pound, J., Mika, P., and Zaragoza, H. Ad-hoc object retrieval in the web of data. In *Proceedings of the 19th International Conference on World Wide Web* (New York, NY, USA, 2010), WWW '10, Association for Computing Machinery, p. 771–780.

[12] Ramsdell, J., and Dietz, L. A large test collection for entity aspect linking. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management* (New York, NY, USA, 2020), CIKM '20, Association for Computing Machinery, p. 3109–3116.

[13] Robertson, S. E., and Jones, K. S. Relevance weighting of search terms. *Journal of the American Society for Information science 27*, 3 (1976), 129–146.

[14] Salton, G., and Buckley, C. Term-weighting approaches in automatic text retrieval. *Information processing and management 24*, 5 (1988), 513–523.

[15] Xiong, C., and Callan, J. Query expansion with freebase. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval* (New York, NY, USA, 2015), ICTIR '15, Association for Computing Machinery, p. 111–120.

[16] Xiong, C., Callan, J., and Liu, T.-Y. Bag-of-entities representation for ranking. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval* (New York, NY, USA, 2016), ICTIR '16, Association for Computing Machinery, p. 181–184.

[17] Xiong, C., Callan, J., and Liu, T.-Y. Word-entity duet representations for document ranking. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2017), SIGIR '17, Association for Computing Machinery, p. 763–772.