

Why does this Entity matter? Support Passage Retrieval for Entity Retrieval

Shubham Chatterjee

sc1242@cs.unh.edu

University of New Hampshire
Durham, New Hampshire

Laura Dietz

dietz@cs.unh.edu

University of New Hampshire
Durham, New Hampshire

ABSTRACT

Our goal is to complement an entity ranking with human-readable explanations of how those retrieved entities are connected to the information need. While related to the problem of support passage retrieval, in this paper, we explore two underutilized indicators of relevance: contextual entities and entity salience. The effectiveness of the indicators are studied within a supervised learning-to-rank framework on a dataset from TREC Complex Answer Retrieval. We find that salience is a useful indicator, but it is often not applicable. In contrast, although performance improvements are obtained by using contextual entities, using contextual words still outperforms contextual entities.

ACM Reference Format:

Shubham Chatterjee and Laura Dietz. 2019. Why does this Entity matter? Support Passage Retrieval for Entity Retrieval. In *The 2019 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '19)*, October 2–5, 2019, Santa Clara, CA, USA. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3341981.3344243>

1 INTRODUCTION

Entity ranking is important for applications which seek either a particular entity or a list of entities, for example, entities involved in the *Brexit* such as *Theresa May*. Several studies show that 40–70% of all web searches target entities [2, 8]. Such queries are best answered by giving the user a ranking of relevant entities. However, in addition to the ranking of entities, presenting a short passage to the user which tells them how or why the entity is related to the query may be more helpful since it helps users decide if they want to know more about the entity. Analogous to search snippets, we want to present a short passage that explains why the entity is relevant to the query. We study the following task:

Task: Given a user’s information need Q ; an external system predicts a ranking of entities E . Our task is to, for every relevant entity $e_i \in E$, retrieve and rank K passages s_{ik} that explain why this entity e_i is relevant for Q —we call these support passages henceforth.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICTIR '19, October 2–5, 2019, Santa Clara, CA, USA

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6881-0/19/10...\$15.00

<https://doi.org/10.1145/3341981.3344243>

As an example, consider the following query-entity pair and the corresponding support passage:

Query: Unfree labour

Entity: Detention (imprisonment)

Support Passage:

Unfree labour is a generic or collective term for those work relations, especially in modern or early modern history, in which people are employed against their will by the threat of destitution, detention, violence (including death), lawful compulsion, or other extreme hardship to themselves or to members of their families.

The novel contribution of this work is a joint model of contextual entity and salience features to rank passages for a given query and entity. We compare our model to two baselines, one baseline retrieves passages for the query-entity pair using a compound query of the query and entity terms, the second baseline scores passages for the query-entity pair by the number of relevant entities (for the query) contained in the passage.

In addition to being relevant for Q , those support passages s_{ik} should mention the entity e_i in a salient way. Salient means that the entity is central to the passage. The salience of an entity can be predicted with the toolkit SWAT [11].

For the query *Brexit* and the entity *Theresa May* consider the following two passages:

Passage 1: If Labour and the Conservatives fail to reach an agreement, MPs will face a series of votes on Brexit options, which could include another referendum. Theresa May has said her government stands ready to abide by the decision of the House if Labour does the same.

Passage 2: British Prime Minister Theresa May offered a new Brexit plan on Tuesday, in a last-ditch effort to get her still-unpopular Brexit deal approved.

While both passages mention *Theresa May* and are about *Brexit*, Passage 2 discusses how Theresa May is affecting Brexit and hence the entity is central to the passage whereas in Passage 1 it is not.

We develop a range of features which incorporate the salience of the entity in the passage for the task of support passage retrieval. We explore if predictions made by SWAT are sufficient to reap the benefits from entity salience. The effectiveness is evaluated in a learning-to-rank [10] setting.¹

¹For more experiments, results, descriptions of methods and datasets, refer to the online appendix for this paper at: <https://shubham526.github.io/support-passage-retrieval/>.

2 RELATED WORK

Sentence Retrieval retrieves relevant sentences in response to a query, question or another sentence. Tasks such as question answering [4], summarization, novelty detection, topic detection and tracking [12], and information provenance make use of a sentence retrieval module as a pre-processing step. However, none of these address the task of support sentence retrieval. The closest is perhaps the work by Cardie et al. [4] which discusses sentence ranking models where the query includes a constraint on a type of entity (e.g. a location, a person). Blanco and Zaragoza [3] present a model that ranks entity support sentences with learning-to-rank. Their work focuses on features based on named entity recognition (NER) in combination with term-based retrieval models.

Voskarides et al. [13] study the problem of explaining relationships between pairs of knowledge graph entities with human-readable descriptions. They model the task as a learning-to-rank problem with a rich set of features which include textual, entity and relationship features. Aggarwal et al. [1] rank all the paths between any two entities in a knowledge graph. This can help in explaining relationships between seemingly unconnected entities.

Kadry and Dietz [9] use relation extraction using OpenIE for support passage retrieval. Their work studies how relation extraction can help in support passage retrieval and the limitations of the current relation extraction approaches that need to be overcome. Their work explores a rich set of features for our task.

None of the works above study joint query-entity and paragraph relevance nor do they consider the effect of entity salience. We propose to look at features that model the joint relevance of a paragraph to an entity and a query, which captures the entity salience in the passage.

3 APPROACH

Given a ranked list of entities for a query, we seek to embellish it with passages which would explain to the user why the entity is relevant to the query. We call the entities in the ranking as target entities. We only try to predict support passages for target entities which are also relevant (according to the entity ground truth for the query). To model this joint relevance for *both* the query and the entity, we use the co-occurring entities with a given entity.

3.1 Candidate Passage Generation

We generate the candidate set of passages using a learning-to-rank combination of the following retrieval models. The RM1/RM3 are expanding with words in this case. Later, we also use RM1/RM3 for expanding the query with entities.

- BM25 (default Lucene) with RM1 (BM25 + RM1)
- BM25 (default Lucene) with RM3 (BM25 + RM3)
- Language Models with Jelinek-Mercer smoothing ($\lambda = 0.4$), with RM1 (LMJM + RM1)
- Language Models with Jelinek-Mercer smoothing ($\lambda = 0.4$), with RM3 (LMJM + RM3)
- Language Models with Dirichlet smoothing, with RM1 (LMDS + RM1)
- Language Models with Dirichlet smoothing, with RM3 (LMDS + RM3)

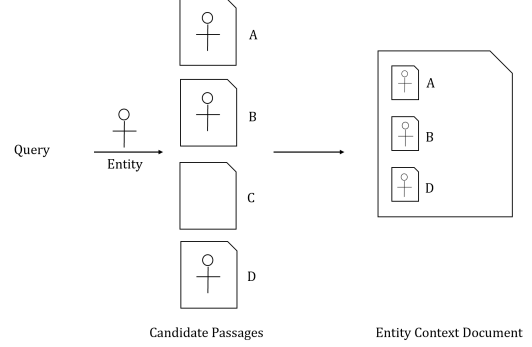


Figure 1: Entity Context Document

3.2 Features

3.2.1 Features based on co-occurring entities. We assume that a passage with a lot of query-relevant entities which co-occur frequently with a given entity is likely talking about that entity. To find most frequently co-occurring entities with a given entity, we filter all passages (retrieved for the query) mentioning the entity and "stitch" them together into D_e which we call an *Entity Context Document* (ECD) about the entity e , following an idea of Dalton et al. [5] (Figure 1). All entities $e' \in D_e$ co-occur with e . A distribution over these co-occurring entities would tell us how likely we are to see the co-occurring entity e' provided we have already seen the query-entity pair (q, e) . We derive this distribution by finding the number of times an entity e' occurs in the ECD. More formally,

$$P(e'|e, q) \propto \sum_{p \in D_e} \text{count}(e' \in p) \quad (1)$$

where p is a paragraph, D_e is the ECD about entity e , $e' (\neq e)$ is an entity co-occurring with e , and $\text{count}(e')$ is the number of times e' is mentioned in p .

- (1) **Entities in the Entity Context Neighbors (ECN).** We score a passage in an ECD D_e by accumulating co-occurrence scores of entities e' contained therein. More formally, we can define a feature value $f_{qe}(p)$ of a paragraph $p \in D_e$ for query-entity pair (q, e) as

$$f_{qe}(p) = \sum_{e' \in p} P(e'|e, q) \quad (2)$$

- (2) **Query expansion using entities (QEE) from ECD.** Using top 20 co-occurring entities to expand the query using RM1 and RM3 on a search index that has a separate field for entities. We experiment with the same variations as given in Section 3.1. Each variation becomes a feature.

3.2.2 Features based on entity context documents.

- (1) **Retrieval score of an ECD.** For every query, we retrieve ECDs using BM25. There is one ECD per entity consisting of passages mentioning the entity. A passage might mention several entities and hence may be present in the ECD of more than one entity. As in Section 3.2.1 (1), we obtain a passage score by accumulating the scores of the ECDs it appears in.

- (2) **Query expansion using words (QEW) from ECD.** Similar to Section 3.2.1 (2), we expand the query using terms from the ECDs using pseudo relevance feedback. We use top 50 terms for expansion and 100 documents as the feedback set. We experiment with the same variations as given in Section 3.1. Each variation becomes a feature.

3.2.3 Features based on entity salience. We use SWAT [11] to annotate passages with salient entities and salience score and class (whether salient or not). We consider $\text{Score}(p|e) = \text{Salience}(e|p)$. We score a passage p for the query-entity pair (q, e) in two ways:

- (1) **Salience score of passage with co-occurring entities (SSP-CE).** We obtain a support passage ranking for every query-entity pair from the method in 3.2.1(1). We re-rank these passages using

$$\text{Score}(p|e, q) = \text{Score}(p|e) \cdot \text{Score}(e|q) \quad (3)$$

where $\text{Score}(e|q)$ is the normalized retrieval score of e (obtained from the entity ranking) and $\text{Score}(p|e)$ is the normalized salience score of e for p .

- (2) **Salience score of passage which mentions an entity (SSP-EM).** We obtain an ECD about an entity and rank the passages in this ECD using

$$\text{Score}(p|e, q) = \text{Salience}(e|p) \cdot \text{Score}(p|q) \quad (4)$$

where $\text{Score}(p|q)$ is the normalized retrieval score of p (obtained from the candidate passage ranking) and $\text{Salience}(e|p)$ is the normalized salience score of e for p .

4 EVALUATION

4.1 Research Questions

We study the extent to which the following components in our approach affect the quality of retrieved support passages:

- RQ1** What is the effect of contextual entities?
- RQ2** What is the effect of entity salience?
- RQ3** Is it better to expand queries with entities or words?

4.2 Evaluation Paradigm

We use the dataset from the TREC Complex Answer Retrieval track [6]² to evaluate our methods. It contains both passage and entity ground truth data as well as an entity linked corpus consisting of paragraphs from the entire English Wikipedia. We use the entity links provided in the collection as well as those annotated using TagMe [7]. We derive a ground truth for entity support passage retrieval from the ground truth of relevant passages and entities provided with the data set (article-level) as follows: any relevant passage that contains an entity link to a relevant entity for the query is defined as relevant for the given query and entity. We apply our methods to produce a passage ranking for every query-entity pair and perform 5-fold cross validation. We use RankLib³ with Coordinate Ascent to optimize for Mean Average Precision (MAP). We use MAP, Mean Reciprocal Rank (MRR), Precision at R (P@R) and Precision at 1 (P@1) as our evaluation metrics—omitted results are available in our online appendix.

²<http://trec-car.cs.unh.edu>

³Dang, V. "The Lemur Project-Wiki-RankLib." Lemur Project,[Online]. Available: <http://sourceforge.net/p/lemur/wiki/RankLib>.

Table 1: Performance with standard error of features individually and combined with L2R, including subsets/ablations.

	MAP	P@1
Baseline 1	0.09±0.002	0.06±0.002
Baseline 2	0.07±0.003	0.06±0.003
ECN	0.31±0.004	0.30±0.004
QEW (LMJM+RM3)	0.35±0.004	0.34±0.004
QEE (LMJM+RM3)	0.30±0.003	0.28±0.004
SSP-CE	0.01±0.002	0.02±0.003
SSP-EM	0.02±0.003	0.03±0.004
L2R-All-Features	0.30±0.004	0.34±0.004
Subset-1 (all – QEW)	0.28±0.004	0.30±0.004
Subset-2 (all – QEE)	0.35±0.004	0.35±0.004
Subset-3 (only QEW)	0.35±0.004	0.34±0.004
Subset-4 (only QEE)	0.30±0.004	0.29±0.004
Subset-5 (QEW + QEE)	0.35±0.004	0.34±0.004

4.3 Baseline

We include two baselines that use query and entity, without any other components of our approach.

4.3.1 Baseline 1: Frequency of relevant entity links. We rank passages for a query-entity pair by the number of relevant entities in the passage. For example, if a passage p contains entities $\{e_1, e_2\}$ and the entities $\{e_1, e_2, e_3, e_4\}$ have been retrieved for the query q , then the score of p for each of the query-entity pairs is $f_{qe_1}(p) = f_{qe_2}(p) = 2$ because the passage has two entities in common with the list retrieved for q .

4.3.2 Baseline 2: Compound entity-query score. We retrieved passages using a compound query, where the query is a combination of the original query and the target entity.

4.4 Results

The most interesting results are presented in Table 1. Due to space constraints other results were moved to the online appendix for this paper. Below, we discuss each of the research questions presented in Section 4.1.

RQ1 (contextual entities): We observe in Table 1 that ranking passages with contextual entities (ECN) achieves a MAP of 0.31 and query expansion with entities (QEE) a MAP of 0.30. This is a significant improvement over the two baselines which have a MAP of 0.09 and 0.07 respectively. This demonstrates the benefit of contextual entities. See Section 4.5 for more discussion.

RQ2 (entity salience): We observe in Table 1 that the two methods which rank passages using entity salience perform the worst. This is also the case when combined with other features using learning-to-rank. We manually confirmed that SWAT correctly identifies salient and non-salient entities. However, only few retrieved entities have a passage with a salient mention. While entities with salient passages are often relevant, a majority (95%) of retrieved entities do not have a passage with a salient mention in the candidate pool. Since the salience feature is only applicable to very few entities, it only has a limited impact on the overall result.

Table 2: Results on subset of entities with at least one salient mention.

	MAP	P@1
Baseline 1	0.16±0.01	0.12±0.01
Baseline 2	0.09±0.01	0.10±0.01
ECN	0.41±0.02	0.37±0.02
SSP-CE	0.47±0.02	0.64±0.02
SSP-EM	0.28±0.01	0.40±0.02

To study whether salience is a useful indicator when it is applicable, we analyze results on the subset of rankings for query-entity pairs for which the passage ranking contains at least one passage in which the target entity is salient. The results are shown in Table 2. We observe that the salience indicator performs very well as compared to other methods. Hence, salience is a useful indicator—however it is only applicable for entities which have a salient passage in the candidate pool. (More results in the online appendix.)

RQ3 (query expansion): We observe that among all methods which retrieve with contextual entities, the method using LMJM with RM3 performs the best with a MAP of 0.30. Note that here we use RM3 to expand the query with entities (see Section 3.2.1(2)). Also, among all methods which retrieve with contextual words, the methods using LMJM with RM3 performs the best with a MAP of 0.35. As we can see, retrieval with contextual words outperforms retrieval with contextual entities and although contextual entities achieve good results, they are unable to outperform contextual words. Hence, contextual words are more informative than contextual entities. Another observation is that among BM25, LMDS and LMJM, retrieval with LMJM always gives the best performance for both contextual entities and contextual words. For more discussion, see Section 4.5.

4.5 Subset Ablation Study

From Table 1 we observe that using QEW features alone obtain a higher MAP score than learning-to-rank with all features. To investigate this further, we conduct a subset/ablation study using learning-to-rank with 5-fold cross-validation using the following combinations:

- (1) **Subset-1:** All features except QE with word features.
- (2) **Subset-2:** All features except QE with entities features.
- (3) **Subset-3:** Only QE with word features.
- (4) **Subset-4:** Only QE with entity features.
- (5) **Subset-5:** Only QE with entity and QE with word features.

The results are shown at the bottom of Table 1.

Regarding RQ1, we observe that although the combination all variations of QEE with entity features (Subset-4) performs well on its own, the ablation study shows that one would obtain even better results by leaving QEE features out (Subset-2) to improve to a MAP 0.35. In contrast, QEW with word features (Subset-3) performs well on their own as well as in the mix with other features as is evident by an increase in MAP of 0.28 on Subset-1 to MAP of 0.30 on

all. Hence, we conclude that contextual entities affect the support passage retrieval to a lesser extent than contextual words.

Regarding RQ3, when L2R is trained to optimize MAP on all contextual word and entity features (Subset-5) the trained parameter assigns the maximum weight to contextual word features (in particular, QE with word using LMJM and RM1/RM3). This shows that contextual words are more informative than contextual entities.

5 CONCLUSION AND FUTURE WORK

This work addresses the task of support passage retrieval to enrich entity rankings in response to a search query. We propose a joint query-entity-passage ranking method and present some initial results. In particular, we show that co-occurring entities are an important indicator of which passages might support an entity for a query. In contrast, retrieving passages using a compound query of the original query and the entity is not sufficient for the problem.

We also experiment with entity salience, and find that it is a highly informative feature when it is applicable, however 95% of target entities do not have any passage with a salient mention in the candidate set. We thus identify a need for developing high-recall salience techniques that are applicable to a larger number of entities, as well as new indexing and retrieval methods that integrate entity salience in an early phase.

REFERENCES

- [1] Nitish Aggarwal, Sumit Bhatia, and Vinith Misra. 2016. Connecting the dots: Explaining relationships between unconnected entities in a knowledge graph. In *European Semantic Web Conference*. Springer, 35–39.
- [2] Krisztian Balog. 2018. *Entity-oriented search*. Springer.
- [3] Roi Blanco and Hugo Zaragoza. 2010. Finding support sentences for entities. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 339–346.
- [4] Claire Cardie, Vincent Ng, David Pierce, and Chris Buckley. 2000. Examining the role of statistical and linguistic knowledge sources in a general-knowledge question-answering system. In *Proceedings of the sixth conference on Applied natural language processing*. Association for Computational Linguistics, 180–187.
- [5] Jeffrey Dalton, Laura Dietz, and James Allan. 2014. Entity query feature expansion using knowledge base links. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 365–374.
- [6] Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. 2017. TREC Complex Answer Retrieval Overview. In *Proceedings of Text REtrieval Conference (TREC)*.
- [7] Paolo Ferragina and Ugo Scaiella. 2010. 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*. ACM, 1625–1628.
- [8] Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 267–274.
- [9] Amina Kadry and Laura Dietz. 2017. Open Relation Extraction for Support Passage Retrieval: Merit and Open Issues. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1149–1152.
- [10] Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval* 3, 3 (2009), 225–331.
- [11] Marco Ponzà, Paolo Ferragina, and Francesco Piccinno. 2018. SWAT: A System for Detecting Salient Wikipedia Entities in Texts. *Computational Intelligence* (04 2018). <https://doi.org/10.1111/coin.12216>
- [12] Nicola Stokes and Joe Carthy. 2001. First story detection using a composite document representation. In *Proceedings of the first international conference on Human language technology research*. Association for Computational Linguistics, 1–8.
- [13] Nikos Voskarides, Edgar Meij, Manos Tsagkias, Maarten De Rijke, and Wouter Weerkamp. 2015. Learning to explain entity relationships in knowledge graphs. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Vol. 1. 564–574.