Conversational Search with Random Walks over Entity Graphs

Gustavo Gonçalves*
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, USA
ggoncalv@cs.cmu.edu

João Magalhães
NOVA LINCS
Universidade NOVA de Lisboa
Lisbon, Portugal
jm.magalhaes@fct.unl.pt

Jamie Callan
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, USA
callan@cs.cmu.edu

ABSTRACT

The entities that emerge during a conversation can be used to model topics, but not all entities are equally useful for this task. Modeling the conversation with entity graphs and predicting each entity's centrality in the conversation provides additional information that improves the retrieval of answer passages for the current question. Experiments show that using random walks to estimate entity centrality on conversation entity graphs improves top precision answer passage ranking over competitive transformer-based baselines.

CCS CONCEPTS

• **Information systems → Information retrieval**; • **Computing methodologies → Discourse, dialogue and pragmatics**.

KEYWORDS

Conversational search, named-entities, entity graph, passage retrieval

ACM Reference Format:
Gustavo Gonçalves, João Magalhães, and Jamie Callan. 2023. Conversational Search with Random Walks over Entity Graphs. In *Proceedings of the 2023 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '23), July 23, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3578337.3605125

1 INTRODUCTION

It is well-established that a query is only an approximate description of an information need. The same is true in conversational search: An individual question only approximates the underlying information need. Typically, other clues are used to infer a better understanding of the information need. Prior research in conversational search uses previous questions or answer passages from the conversation to augment understanding of the current query [23].

We observe that conversational information seeking often explores topics related to named entities, for example, *the Grateful Dead*, *oat milk*, and *bees* (all TREC Conversational Assistance Track (CAsT) topics [6, 7]). We hypothesize that *modeling the conversational turn as an entity graph* may be effective, because the most likely interpretations of the current question will be closely connected to entities in candidate answers. Forming an entity graph from multiple sources of evidence creates a more informative representation of the current conversational context.

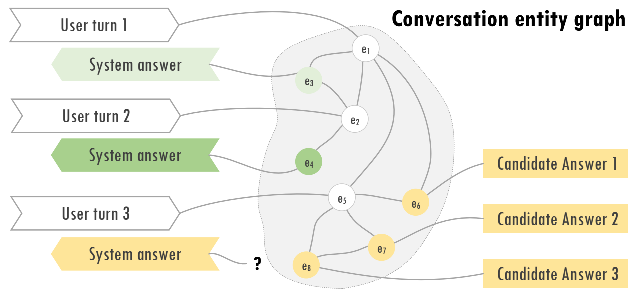

Figure 1: Conversation Entity Graph.

As illustrated in Figure 1, such an entity graph covers the entities mentioned in the current question and retrieved answers. Some of the entities obtained from the retrieved answers are reasonable facets of the current topic. Some are only peripherally related to the focus of the conversation. The challenge in using such a graph is to distinguish the most central or important entities, and use them to improve the understanding of the current question [38, 45].

The inspiration behind this work is related to the idea that entities of a document ranking, can be used to reach similar documents given the connections of an entity graph built with the top documents of the document ranking [30]. Under a conversational scenario, the interaction between the query and highest-ranked passage candidates can be insufficient to cover different conversational facets. While previous approaches [23] have successfully modeled near-context across turns using queries, they ignore entity interactions between passages.

This paper proposes a novel approach that is added to a standard transformer reranking architecture. It uses the top-ranked passages of an initial retrieval to form an entity graph that represents the current conversation turn, estimates the importance or centrality of each entity to the turn, and uses these estimates to rerank the retrieved passages. Experimental evaluation shows that the method improves precision at the top of the rankings for TREC CAsT datasets, which is ideal for conversational assistants where only a few answers are required.[1] The reranking method's improvements, albeit modest, are offset by its low computational cost, making the method an attractive addition to a conversational system. Finally, we study the influence of the entity graph design for conversational search.

The next section discusses published research related to this work. Section 3 describes the formation of entity graphs and computation of entity centrality scores. Section 4 discusses the use of

*Also with Universidade NOVA de Lisboa, NOVA LINCS.

[1]Source code: https://github.com/gsgoncalves/ICTIR2023-ConvSearchWithEntGraphs

ICTIR '23, July 23, 2023, Taipei, Taiwan

ACM ISBN 979-8-4007-0073-6/23/07.
https://doi.org/10.1145/3578337.3605125

those scores in passage ranking. Sections 5 and 6 present the experimental methodology and experimental results. Section 7 concludes.

2 RELATED WORK

Conversational search introduces novel dimensions to the ad-hoc retrieval scenario going beyond the traditional list of search results. Previous works have already shown that explicitly tackling named-entities can improve many language modeling tasks [15, 22]. Neural architectures still present room for improvement, as they do not fully discriminate the important textual information available [35]. We hypothesize that named-entities are a possible information source to bridge the semantic gaps introduced by pre-trained language models to rerank passages in a conversational search task.

There are two challenging main aspects that distinguish conversational search from ad-hoc ranking tasks: First, it focuses on a live dialogue scenario emphasizing the importance of the top-1 results; Second, it adds a challenging dimension of sequence between utterances, thus introducing a notion of context, or history.

Task-based conversational search tasks illustrate both previous aspects by focusing on unforgiving scenarios that prioritize the quality of the top results, while maintaining conversational context [13, 36]. Our work focuses mainly on the first scenario, where passage ranking tasks have been adapted to include conversational context. Popular approaches to passage ranking, while maintaining a conversational context, include using the transformer architecture to encode context along with information needs [12, 19, 31, 32]. These approaches leverage on pre-trained language models to obtain semantic context they are limited by the amount of information they can encode, since they are very memory intensive.

Given the limited information that novel strategies can encode it is also necessary to build better queries. Query rewriting is a competitive strategy in approximating the users' information need as the conversation evolves, and has been largely studied to satisfy users' information needs [23, 24, 37]. Additionally, it has been shown that using graph structures for multi-hop strategies is a competitive approach to estimate important terms during a conversation [5, 46, 48]. Graph-based methods allow the exploration of neighboring levels of a knowledge base, which can be used to infer the topics a conversation might follow.

External knowledge bases provide additional information that may not be explicit in documents. A fundamental aspect is the extraction and linking of named-entities across the conversation turns. Before conversational search took centre stage, many entity linking works were proposed [11]. One family of approaches uses some form of external knowledge such as Wikipedia or DBpedia [1]. TagMe [11] and DBPediaSpotlight [25], are long-standing examples of such approaches. More recently, other approaches extend the external knowledge with representation learning in the form of embeddings such as Wikipedia2Vec [42]. Good examples include REL [39], BLINK [21], and GENRE [8].

Driven by the research in conversational assistants, Joko et al. [17] examined how different entity linkers behaved in this domain. The authors observed that deep learning methods achieve a higher precision but very low recall. Overall, the best f-measure was achieved by the methods based on textual representations of Wikipedia [11].

There is a wide range of works exploring named-entities with the Transformer architecture [29, 47]. While these works have been successful in a number of tasks, there is not enough evidence that such approaches can improve the Transformer architecture in ad-hoc retrieval tasks, or conversational search tasks. It is interesting to see that previous improvements with named-entities [40, 41] have yet to be translated into the Transformer generation of ranking methods. See the experiments section of this paper for more details.

3 CONVERSATION ENTITY GRAPHS

Our goal is to improve the top precision of conversational search, as this is an important factor for user satisfaction when dealing with conversational assistants. We hypothesize that the entities appearing in top-ranked passages are connected both by their co-occurrence and semantic relations, which can provide access to an extended set of relevant passages at lower-ranked positions that might be overlooked by neural rankers. We propose a lightweight approach that leverages on the top passage results of state-of-the-art neural rankers, and encodes the query-passage interactions across the ranking as an entity graph.

Modern state-of-the-art ranking pipelines often start with a lexical ranker, such as BM25 [33], to quickly obtain a set of documents that approximate a user's information need. Increasingly computationally expensive rerankers are applied subsequently to reorder the documents and maximize relevance. Recent work uses neural language models, such as BERT, to reorder the documents obtained by the earlier ranker(s) [14].

This work proposes a lightweight reranking model that utilizes entity graphs as a representation of conversational context, and consists of two distinct stages. The first stage is a *full-text retrieval ranker*. In our work, the full-text retrieval ranker consists of a query likelihood ranker followed by a BERT [9] reranker, but any full-text retrieval system may be used. The full-text retrieval ranker produces a ranking of passages $P = \{p^1, \cdots, p^k\}$, an ordered collection of k candidate passages that answer a query q. The second stage analyzes P to estimate *entity centrality* scores, which are used to rerank passages more effectively. We refer to these two stages as the *full-text retrieval* and *entity centrality* stages throughout this paper. The following subsections explain the *entity centrality* stage of the reranking model, by decomposing the graph construction in Section 3.1, defining how to weight the graph edges in Section 3.2, and finally how to determine entity centrality in Section 3.3.

3.1 Nodes and Edges of the Entity Graph

This work, inspired by previous competitive approaches that consider named-entities as a connective element between documents [10, 40, 41], infers the current conversation context by estimating centrality over an entity graph, thus connecting query and passages.

Named-entities in queries and passages provide a knowledge-aware view of the textual content. Linking text to a knowledge-base is a starting point to obtain external connections that are not explicit in the query-passage text. We argue that entity occurrence is enough to provide information and reweight the full-text retrieval ranker passage scores. Thus, giving more importance to passages that contain entities central to the current query, but lack the exact query terms to be highly ranked by the full-text retrieval ranker.

The turn-specific entity graph construction begins with linking the entities [17] of the query and respective retrieved passages obtained by the full-text retrieval ranker. The nodes of the graph are given by the entities in the passages, and the edges correspond to the occurrence in the passages. The set of unique entities E is computed from the current conversation query q, and the top retrieved passages P. This leads to the set of unique entities E defined as:

$$E = \{e_1, \cdots, e_g, \cdots, e_n\}, \forall e_g \in \{q\} \cup P \tag{1}$$

Given the set of n entities, E, and the top k passages P, we compute the entities-passage occurrence matrix $\mathbf{C_P}$ as:

$$\mathbf{C_P} \in \{w_g\}^{n \times k} \tag{2}$$

To build the affinity matrix we consider the weighted occurrences of entities $\mathbf{C_P}$ for each query q of the conversational search session, which will result in the entities that will be in the graph to calculate the centrality scores and rerank the top passages. The query vector $\mathbf{C_Q}$, Eq. (3), contains the entities mentioned in the query and we define it as a multi-hot vector:

$$\mathbf{C_Q} \in \{0, 1\}^{n \times 1} \tag{3}$$

This allows us to compute the occurrence matrix of the conversation entities $\mathbf{C_{QP}}$, as the concatenation of vector $\mathbf{C_Q}$ with matrix $\mathbf{C_P}$,

$$\mathbf{C_{QP}} = \left[\ \gamma \cdot \mathbf{C_Q}\ \Big|\Big|\ (1-\gamma) \cdot \mathbf{C_P}\ \right]^T. \tag{4}$$

The choice of concatenating the query vector with the passage entity matrix was motivated by the idea that relevant documents would contain similar entities contained in the query, thus it is important that the query also makes part of the entity matrix.

Eq. (4) introduces a linear combination parameter γ. The γ parameter controls the weight balance between the entities in the query vector and the entities in the top passages matrix. Motivated by linear interpolation schemes such as the Jelinek and Mercer [16] smoothing, this parameter allows flexibility to weight the query and passages entities differently and observe how their contribution affects the results. In our experimental results we tune γ to obtain the optimal weight to be given to the entities in the query vector, or the top passages matrix. By expanding Equation (4) we get the unrolled expression:

$$\mathbf{C_{QP}} = \left[\ \gamma \cdot \begin{bmatrix} q_{e_1} \\ \vdots \\ q_{e_n} \end{bmatrix} \ \Bigg|\Bigg|\ (1-\gamma) \cdot \begin{bmatrix} p^1_{e_1} & \cdots & p^k_{e_1} \\ \vdots & \ddots & \vdots \\ p^1_{e_n} & \cdots & p^k_{e_n} \end{bmatrix} \ \right]^T. \tag{5}$$

Finally, the entity graph, Eq. (6), is given by the application of the dot product over the occurrence matrix

$$\mathbf{G} = \mathbf{C_{QP}} \cdot \mathbf{C_{QP}}^T, \qquad \mathbf{G} \in \mathbb{R}^{n \times n}. \tag{6}$$

3.2 Weighting the Entity Graph Edges

The weighting scheme used in the previous subsection is obtained by signaling the presence of entities in passages, and query. A more informative, conversation-specific, weighting scheme can be further designed with the passage rank scores. In this weighting scheme, the values of $\mathbf{C_P}$ correspond to the full-text retrieval ranker scores, RS. Hence, the score of each passage entity of $\mathbf{C_P}$ is given by:

$$cp_e = RS(p_e) \quad \forall e \in E \wedge \forall p \in P \tag{7}$$

Using the Equation (7) weights in Equation (6) is equivalent to setting all the values of each column to the full-text retrieval ranker score of the corresponding passage. The weight given to the entities in the query vector $\mathbf{C_Q}$ is the original multi-hot binary encoding.

The model uses this graph edge weighting scheme to maintain a strong signal from the query entities. Moreover, this formulation allows for entity occurrences in higher-ranked passages to have more influence than entity occurrences in lower-ranked passages.

3.3 Calculating the Entity Graph Centrality

The entity graph represents the entities related to the current conversation turn and how they are used in the query and top-ranked passages. The next step is to calculate the entity centrality (EC) scores that indicate how well each entity represents the conversation turn context.

The EC scores can be estimated with random walk methods. We focus on eigenvector methods as an implementation of random walks to estimate centrality [2, 4], as they can be implemented efficiently through a power-iteration with convergence in $O(Edges \times Iterations)$. Moreover, we choose a particular use case of the eigenvector centrality [3] with a teleportation variation.

The **EC** vector of the top passages entities is computed as

$$\mathbf{EC}^{(t)} = (1-\alpha) \cdot \frac{1}{|E|} + \alpha \cdot \mathbf{G} \cdot \mathbf{EC}^{(t-1)} \tag{8}$$

where α is the damping factor and each dimension i of **EC** contains the centrality score of entity i.

Over both datasets the best results were achieved by setting the dampening factor to 0.99, virtually eliminating the teleportation factor introduced by Eq. (8), as our task relies on small connected graphs that require a small amount of dampening. We keep the dampening factor, ever so slightly, for the sake of guaranteed convergence of the power-iteration algorithm [20].

4 RERANKING WITH ENTITY CENTRALITIES

Formally, the score of each top passage is obtained by computing the dot product between **EC** scores, and the entity-passage matrix $\mathbf{C_P}$.

$$\mathbf{S} = \mathbf{EC}^T \cdot \mathbf{C_P}, \qquad \mathbf{S} \in [0,1]^{1 \times k} \tag{9}$$

Eq. (9) results in a scoring vector for all of the full-text retrieval passages in matrix $\mathbf{C_P}$, now conditioned on entity information.

With the score vector defined in Eq. (9) we can perform a reranking step based on the entity centralities of the passage. We refer to this scoring system as EC_{binary}.

A straightforward extension to EC_{binary} is to fuse the entity centrality ranking, with the full-text retrieval ranking. Motivated by Jelinek and Mercer [16], we balance the original scores derived from the full-text retrieval ranker with the entity centrality scores. The linear interpolation scoring is formalized below for any passage k in the ranking:

$$p^k = (1-\delta) \cdot \mathrm{S}^k + \delta \cdot RS^k \tag{10}$$

Table 1: Retrieval baselines compared with the averages for the 5-Fold CV Entity Centrality re-ranking. Statistically significant improvements are denoted with †, and non-inferiority with *, for $p < 0.05$ with a margin of 0.01, over the BERT baseline.

	CAsT 2019					CAsT 2020				
Method	**nDCG@1**	**nDCG@3**	**P@1**	**P@3**	**MRR**	**nDCG@1**	**nDCG@3**	**P@1**	**P@3**	**MRR**
Term based approaches										
BM25	0.4152	0.3858	0.6012	0.5568	0.7157	0.2528	0.2536	0.3798	0.3798	0.5241
LMD	0.3974	0.4026	0.5838	0.5896	0.6984	0.3257	0.2930	0.4952	0.4167	0.6024
RM3	0.4099	0.4133	0.6069	0.6031	0.7158	0.3013	0.2808	0.4519	0.4135	0.5690
BERT	0.5689	0.5703	0.7803	0.7476	0.8604	**0.5244**	0.4976	0.6923	0.6538	**0.7783**
Entity based approaches										
ERNIE	0.5626	0.5617	0.7514	0.7245	0.8435	0.5243	0.4865	**0.6971**	0.6394	0.7750
E-BERT	0.5270	0.5205	0.7283	0.6802	0.8229	0.4006	0.3786	0.5673	0.5208	0.6840
EC_{binary}	0.6074	0.5839*	0.8035	0.7534	0.8707	0.4812	0.4950	0.6635	0.6554	0.7598
EC_{BERT}	0.6320†	**0.6164**†	**0.8439**†	**0.7746***	0.8869*	0.5088	**0.5092***	0.6779	**0.6779***	0.7713
EC_{linear}	**0.6334**†	0.6102†	**0.8439**†	0.7649	**0.8871***	0.5104	0.5084*	0.6779	0.6731*	0.7730

The balance between the passage centrality score, S^k, and the full-text retrieval ranker, RS^k, score is tuned with the hyperparameter δ. The motivation for combining S^k and the ranking provided by the full-text retrieval ranker is to retain the full-text retrieval score since it captures complementary relevance signals, including interactions among query and passage terms that do not correspond to entities. We name this scoring system as EC_{linear}.

5 EXPERIMENTAL METHODOLOGY

Datasets: The TREC CAsT [6, 7] benchmark provides evaluation datasets for conversational search. It is composed of the MSMarco [27] and the TREC CAR Wikipedia datasets [26]. The CAsT datasets follow a dialog construction, where the last utterances of a dialog combine information needs that have occurred during the conversation. We use the set of manual queries for the 2019 and 2020 editions of the dataset to maximize entity recall.

Entity Linking: Entity Linking is a preprocessing step that can be performed offline for corpora, and at runtime for the queries. We opted to use TagMe [11] as the entity linker for its superior F-measure on the CAsT 2019 and 2020 datasets [17]. TagMe used a Wikipedia dump from November 2019 as its knowledge base, and we linked entities with a confidence score of 0.1 to maximize entity recall on both queries and passages.

Baselines: We compare the proposed methods with three classical retrieval models and three transformer models. The classical retrieval models are BM25 ($k = 1.1$, $b = 0.3$), LMD ($\mu = 1000$), RM3 over the previous LMD baseline (5 terms, 15 docs, query weight of 0.8). A BERT reranker is the main baseline, and the starting ranking for entity centrality methods. The BERT reranker was obtained from the LMD run listed in Table 1 and was finetuned [28] on the MSMarco dataset [27], (sample size=100k steps; learning rate=3×10^{-6}; warm-up=10%; ADAM [18] β_1=0.9, β_2=0.999; L2 decay=0.01). We applied the same fine-tuning process to train two other entity-aware transformer models, ERNIE [47] and E-BERT [29].

Finally, the three experimental systems are EC_{binary}, EC_{BERT}, and EC_{linear}. EC_{binary} uses the binary co-occurrence matrix to calculate the entities' centrality. EC_{BERT} replaces the non-zero positions of the matrix with the respective BERT passage score for all entities contained in that passage, before calculating the centrality scores. In both EC_{binary} and EC_{BERT}, the score of each passage is the sum of all entity centrality values. EC_{linear} builds on EC_{BERT} and calculates the passage score as the linear interpolation between, the BERT query-passage score and centrality score. The EC variations are reported with a 5-fold cross-validation over the CAsT 2019 and 2020 datasets.

Evaluation Metrics: The goal of conversational search is to answer a question with the top passage, thus we focused on Precision at ranks 1 and 3. We also measured results with MRR, and nDCG at 1 and 3 to account for the multi-level relevance judgments.

6 RESULTS AND DISCUSSION

This section discusses experimental results and the impact of the system components on the conversational search task.

6.1 Analysis of Top Retrieved Passages

Table 1 shows the retrieval results for all methods. As expected, BERT outperforms the traditional rankers across all metrics. Interestingly, the neural entity-based approaches fall behind BERT, despite being trained in the same way. These neural entity-based architectures learn a deep contextual representation by fusing entity embeddings in the case of ERNIE [47] or transposing entity embeddings to a BERT-compatible embedding space as in the case of E-BERT [29]. However, our experiments show that the additional contextual entity representation diminishes the ranking capabilities of the pre-trained language models.

The centrality-based approaches using a graph built with the top 20 passages show the benefit of using the entities of lower-ranked passages to improve the quality of the top positions of the ranking. Our experiments show gains in combining the Entity Centrality (EC) information with the original BERT ranking.

Statistical significance was determined using two-sided paired t-tests, and non-inferiority with one-sided paired t-tests, following Sakai [34]. The multiple tests were adjusted with the Holm-Bonferroni correction. For both datasets, the EC_{BERT} and EC_{linear}

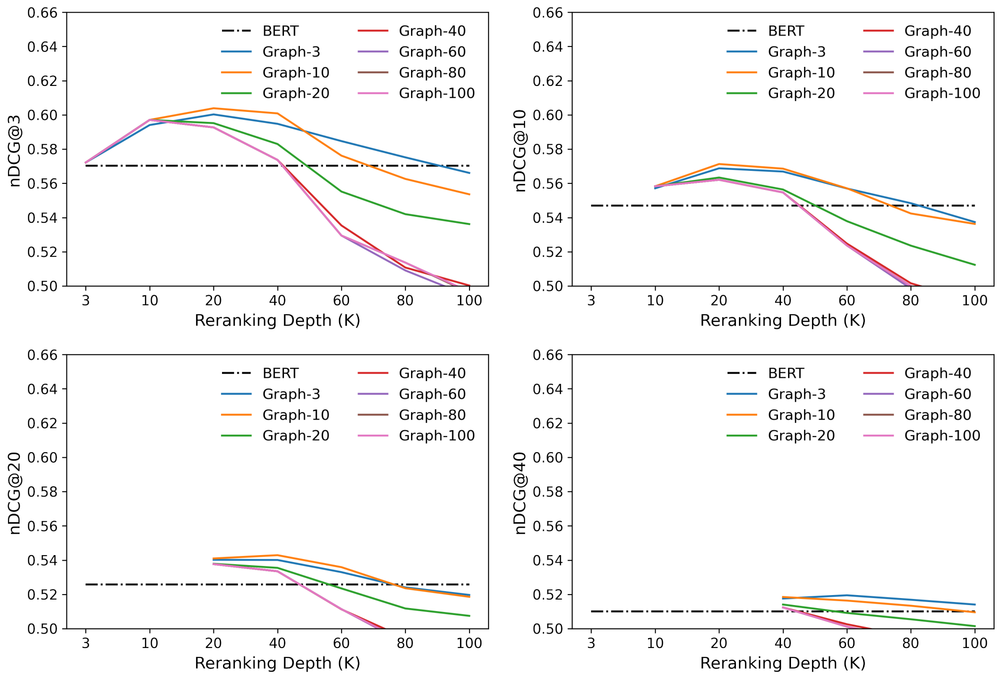

Figure 2: nDCG@3, 10, 20, and 40 after reranking the top K passages on CAsT 2019. Graph-* shows the graph size with the entities from the specified number of passages.

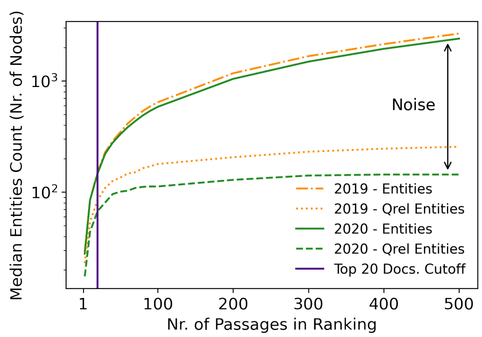

Figure 3: Entity Graph size vs. Number of passages

methods show improvements over the baseline. For 2019 we can observe that nDCG@3 statistically outperforms BERT with a p-value inferior to 0.05, and a relative improvement of 8.1%. nDCG@1 and P@1 are statistically superior to BERT with a p-value inferior to 0.05, and with relative improvements of 11.3% and 8.2% respectively. For the 2020 dataset, EC_{BERT} and EC_{linear} are statistically equal or superior to BERT with a p-value inferior to 0.05 for metrics nDCG@3 and P@3, with relative improvements of 2.3% and 3.7%.

The results show a more modest improvement from 2019 to 2020. Our experiments showed that for the 2020 dataset, on average lower δ provided the best results thus giving less emphasis to the query entities, which means that the best 2020 results were achieved with a lower contribution of the query entities. This behavior is surprising as query entities are the main signal for relevance. It suggests that the 2020 dataset contains a noisier set of entities, which is directly linked to the quality of the contextual entity graphs as discussed in Section 6.2. We found a higher dissociation between the presence of query entities in relevant passages from 2019 to 2020, with 78% of relevant passages containing at least 1 query entity in 2019, and 66% in 2020. Furthermore, for the 2019 edition, only 3.41% of turns do not contain any query entity, when compared to the 7.41% of turns without entities for 2020. These results indicate that the method has fewer connections available to reach relevant passages for the 2020 dataset.

6.2 Entity Graphs over Conversation Turns

The next experiment examines the impacts of the entities' graph quality and the value of the entities added to the centrality-based reranking. Figure 3 shows the divergence between the median set of relevant entities and total entities in the graph across all queries, as we consider more passages of the ranking. After 20 passages (purple vertical line) we start to see a significant increase in noisy entities, i.e. the gap between the entities that occur in relevant passages and entities from all passages. This pattern seems to be linked to the performance difference between the 2019 and the 2020 results that we presented in the previous section: the quality of the entity graph had a positive impact in the 2019 dataset, while in the 2020 dataset, the conversational entity graph is noisier, thus resulting in smaller improvements.

Next, the retrieval performance is studied conditioned on the graph size. The centrality reranking approach has two hyperparameters. The first hyperparameter controls the number of passages from which entities are extracted to build the graph. The second

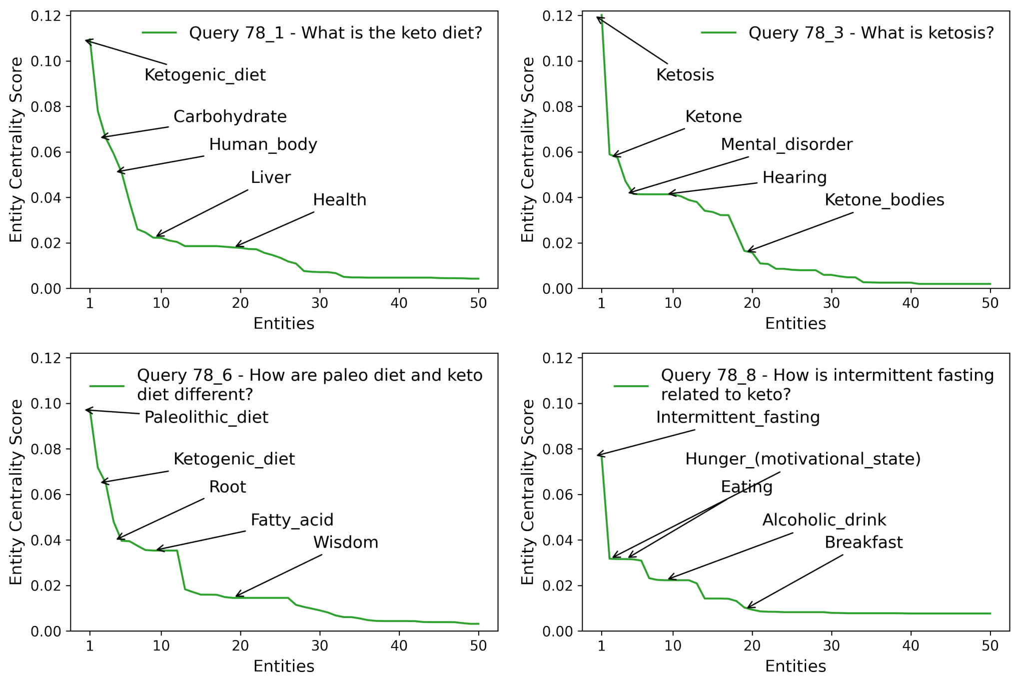

Figure 4: Evolution of the top entities along the conversation topic 78 - 'diet information'. The line plots show the most central entities at rank positions 1, 3, 5, 10, 20.

hyperparameter is the number of passages reranked by the centrality measure, limiting passages to be reordered with the information present in the entity graph. Figure 2 shows the nDCG@3, 10, 20, and 40 for rerankers with different entity graph sizes, on CAsT 2019. Please note that for each nDCG@k graph, the reranking depth (x axis) begins at k to align the metric with the ranking size. The dotted black line shows the BERT nDCG at the corresponding nDCG cutoff. As shown across the four nDCG cutoffs, using entity information from passages further down in the ranking helps to rerank the BERT ranking, which indicates that the method is capturing ranking signals that were not considered by the BERT ranker.

Figure 3 and Figure 2 are connected by the graph depth, and consequently, the entities that are used to estimate the centrality and rerank the passages. Figure 2 shows the different retrieval performances of the EC_{Linear} system, for CAsT 2019, as the entity-graph size increases. Each line in Figure 2 corresponds to a different number of passages used to build the graph e.g., the line "Graph-20" reorders the BERT run using the entities contained in the top 20 passages. Figure 3 shows that on median 20 passages will provide a graph with approximately 150 entities.

Noting that the black dashed line of Figure 2 represents the BERT baseline we can observe that the systems with Graph-10, Graph-20, and Graph-40 beat the baseline at reranking cutoffs up to 40. This confirms the previously seen low divergence for 2019, to the left of the vertical line of Figure 3, between the relevant entity set and the retrieved entity set. To the right of the vertical line of Figure 3 as the divergence between sets increases, the EC model performance also decreases across all systems that use more than 40 passages to build the graph. After this point, the introduced noisy entities lead to sharp drops in performance across all graph sizes, showing the need for a balance between graph size, and ranking depth. Large graphs with many non-relevant entities for the passages to be reranked, or on the other hand, small graphs that do not cover the relevant entities of the passages to be reranked will lead to deficient results.

Another interesting observation, that confirms the observations so far, is that as the graph size increases, there are faster diminishing returns as the reranking depth is also increased. That is, for a graph built with 100 passages (Graph-100 - pink line in Fig. 2), there is too much noise in the graph to rerank more than 40 passages, thus the centrality ranking signals perform worse than the BERT baselines. This observation ties back to Figure 3, where the median graph size built with 100 passages has approximately 500 entities (in Figure 3 when Nr. of Passages in Ranking = 100, the Median Entity Count is ≈ 500). Many of these entities will be noise as we can see from the gap between the lines in Figure 3.

6.3 Qualitative Analysis of Conversation's Rank of Entities

Figure 4 examines the quality of the entities obtained by the random walks for four questions in CAsT conversation 78 to investigate how centrality changes throughout a conversation. It shows the entities ranked at position 1, 3, 5, 10, and 20 by their entity centrality score across the conversation turns 1, 3, 6 and 8.

As the conversation advances from turns 1 and 3, to turns 3 and 6, the entity with the highest score "Ketogenic Diet" and "Ketosis", gives place to subtopics of the conversation on "Paleolithic Diet" and "Intermittent Fasting". It is noteworthy that the first entities

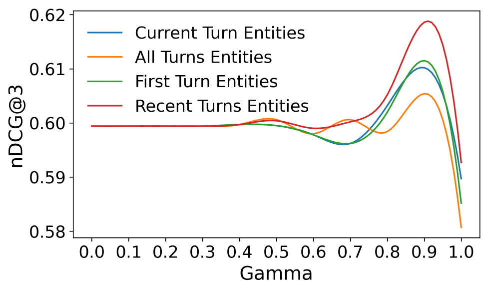

Figure 5: The modes of carrying context on the query side as a function of γ for the CAsT 2019 dataset.

with the highest entity centrality scores are query entities. However, the immediate entities in position 3 are closely related to the query entities and the initial intent, which expands the vocabulary being considered and brings passages that contain these closely connected entities to higher positions in the ranking.

Entities at positions 10 and 20 become less related to the topic turn. This shows that a majority of the centrality score is strongly focused on query entities and their connections, thus making the centrality model grounded in the query entities and closely related entity neighbors.

6.4 The Role of Query Entities

In this section we expand our analysis by doing preliminary work on using different combinations of query entities to maintain conversational context on CAsT 2019. In particular, we examine the impact of enriching the initial query vector C_{Q_j} with entities from previous turns to better capture the conversation context [23, 43, 44]. We hypothesize that maintaining query entities that were mentioned in previous utterances creates a high-level context representation that can roughly approximate the conversation history. By using different strategies to capture query entities we can improve the robustness of the entity graph against topic shifts, while keeping an evolving context of the conversation given by the central entities. We defined the following modes of carrying context as the system advances through conversation turns as follow:

- **Current Turn Entities:** C_{Q_j} contains the entities of the current query.
- **All Turns Entities:** C_{Q_j} contains all entities from previously seen queries, including the current query.
- **First+Current Turn Entities:** C_{Q_j} contains the entities of the first query and the current query.
- **Recent+Current Turn Entities:** C_{Q_j} contains the entities of the three previous queries and the current query.

In Figure 5 and Table 2 we analyze the effects of four query entity combinations while varying the importance of the query entities on the entity graph. The "Current Turn Entities" experimental system sets the lower bound for manipulating conversational context using entities. In this system, no information is carried between conversation turns. The "All Turns Entities" baseline carries all entities on the query side along the conversation, which causes a loss across all cutoffs of nDCG shown in Table 2. A conversation

Table 2: Conversational context combinations on the query side for the CAsT 2019 dataset at γ=0.9.

Graph Query Mode	nDCG@3	nDCG@10	nDCG@20
Current Turn Ent.	0.612	0.568	0.536
All Turns Ent.	0.602	0.560	0.533
First Turn Ent.	0.610	0.569	0.537
Recent Turns Ent.	**0.619**	**0.571**	**0.538**

can have similar information needs that might change the higher-level context of the conversation, which requires the system to give less importance to entities that are no longer central to the current stage of the conversation. Hence, maintaining entities that appeared early in the conversation can harm the results of the final utterances of the conversation. Finally, the most competitive approaches are either adding the entities that appear in the first query – "First Turn Entities", or using the entities of the previously three seen queries – "Recent Turns Entities". Using a recent conversation history, consistently outperforms the remaining combinations. We can infer that entities that appeared closely in previous utterances are related to the current query.

We must note that the improvements across these different combination modes are in very close proximity to each other. This is an opportunity and tentative path to explore to improve results in this conversational search scenario.

7 CONCLUSIONS AND FUTURE WORK

This paper proposes an Entity Centrality method for improving top-3 passage ranking in conversational search. A conversation turn-specific graph is built using the entities from both queries and passages given by any neural ranker. At runtime, random walks are used to estimate the entity centralities over the conversation graph and used to rerank the passages. Experiments demonstrate an improvement of up to 8.1% in nDCG@3 and 3.6% in P@3 on the CAsT 2019 dataset. Results on CAsT 2020 were less competitive and illustrate the importance of having a sufficiently large number of relevant entities in the top passages. In fact, our analysis showed that queries are the main source of relevant entities that approximate closely related entities in passages. Passages are extremely entity-rich, introducing many non-relevant entities in the entity graph, thus the query entities are a strong signal to keep the graph on topic.

ACKNOWLEDGMENTS

This work has been partially funded by the FCT project NOVA LINCS Ref. UIDP/04516/2020, by the Amazon Science - TaskBot Prize Challenge and the CMU|Portugal projects iFetch (LISBOA-01-0247-FEDER-045920) and GoLocal (CMUP-ERI/TIC/0046/2014), and by the FCT Ph.D. scholarship grant SFRH/BD/140924/2018. Any opinions, findings, and conclusions in this paper are the authors' and do not necessarily reflect those of the sponsors.

REFERENCES

[1] Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, Philippe Cudré-Mauroux, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, and Gerhard Weikum (Eds.). 2007. *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings.* Lecture Notes in Computer Science, Vol. 4825. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-76298-0

[2] Michele Benzi and Christine Klymko. 2015. On the Limiting Behavior of Parameter-Dependent Network Centrality Measures. *SIAM J. Matrix Anal. Appl.* 36, 2 (Jan. 2015), 686–706. https://doi.org/10.1137/130950550

[3] Sergey Brin and Lawrence Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems* 30, 1-7 (April 1998), 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X

[4] Seungjin Choi. 2005. On Variations of Power Iteration. In *Artificial Neural Networks: Formal Models and Their Applications - ICANN 2005, 15th International Conference, Warsaw, Poland, September 11-15, 2005, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 3697)*, Wlodzislaw Duch, Janusz Kacprzyk, Erkki Oja, and Slawomir Zadrozny (Eds.). Springer, 145–150. https://doi.org/10.1007/11550907_24

[5] Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abujabal, Jyotsna Singh, and Gerhard Weikum. 2019. Look before You Hop: Conversational Question Answering over Knowledge Graphs Using Judicious Context Expansion. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu (Eds.). ACM, 729–738. https://doi.org/10.1145/3357384.3358016

[6] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2019. CAsT 2019: The Conversational Assistance Track Overview. In *Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, November 13-15, 2019 (NIST Special Publication, Vol. 1250)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST).

[7] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. CAsT 2020: The Conversational Assistance Track Overview. In *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020 (NIST Special Publication, Vol. 1266)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST).

[8] Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2021. Multilingual Autoregressive Entity Linking.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[10] Bhuwan Dhingra, Manzil Zaheer, Vidhisha Balachandran, Graham Neubig, Ruslan Salakhutdinov, and William W. Cohen. 2020. Differentiable Reasoning over a Virtual Knowledge Base. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.* OpenReview.net.

[11] Paolo Ferragina and Ugo Scaiella. 2012. Fast and Accurate Annotation of Short Texts with Wikipedia Pages. *IEEE Software* 29, 1 (2012), 70–75. https://doi.org/10.1109/MS.2011.122

[12] Rafael Ferreira, Mariana Leite, David Semedo, and João Magalhães. 2021. Open-Domain Conversational Search Assistant with Transformers. In *ECIR (1) (Lecture Notes in Computer Science, Vol. 12656)*. Springer, 130–145.

[13] Rafael Ferreira, Diogo Silva, Diogo Tavares, Frederico Vicente, Mariana Bonito, Gustavo Gonçalves, Rui Margarido, Paula Figueiredo, Helder Rodrigues, David Semedo, and João Magalhães. 2022. TWIZ: The Multimodal Conversational Task Wizard. In *ACM Multimedia*. ACM, 6997–6999.

[14] Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. Rethink Training of BERT Rerankers in Multi-stage Retrieval Pipeline. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12657)*, Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). Springer, 280–286. https://doi.org/10.1007/978-3-030-72240-1_26

[15] Hao Huang, Xiubo Geng, Jian Pei, Guodong Long, and Daxin Jiang. 2021. Reasoning over Entity-Action-Location Graph for Procedural Text Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 5100–5109. https://doi.org/10.18653/v1/2021.acl-long.396

[16] F Jelinek and Robert Mercer. 1980. Interpolated Estimation of Markov Source Parameters from Sparse Data. *Pattern recognition in practice. Proc. workshop Amsterdam, May 1980* (1980), 381–397, 401.

[17] Hideaki Joko, Faegheh Hasibi, Krisztian Balog, and Arjen P de Vries. 2021. Conversational Entity Linking: Problem Definition and Datasets. *arXiv preprint arXiv:2105.04903* (2021). arXiv:2105.04903

[18] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.

[19] Vaibhav Kumar and Jamie Callan. 2020. Making Information Seeking Easier: An Improved Pipeline for Conversational Search. In *EMNLP (Findings) (Findings of ACL, Vol. EMNLP 2020)*. Association for Computational Linguistics, 3971–3980.

[20] Amy Nicole Langville and Carl Dean Meyer. 2003. Survey: Deeper Inside PageRank. *Internet Math.* 1, 3 (2003), 335–380. https://doi.org/10.1080/15427951.2004.10129091

[21] Belinda Z. Li, Sewon Min, Srinivasan Iyer, Yashar Mehdad, and Wen-tau Yih. 2020. Efficient One-Pass End-to-End Entity Linking for Questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6433–6441. https://doi.org/10.18653/v1/2020.emnlp-main.522

[22] Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-Relation Extraction as Multi-Turn Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1340–1350. https://doi.org/10.18653/v1/P19-1129

[23] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2020. Query Reformulation Using Query History for Passage Retrieval in Conversational Search. *CoRR* abs/2005.02230 (2020). arXiv:2005.02230

[24] Zhongkun Liu, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Maarten de Rijke, and Ming Zhou. 2021. Learning to Ask Conversational Questions by Optimizing Levenshtein Distance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 5638–5650. https://doi.org/10.18653/v1/2021.acl-long.438

[25] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia Spotlight: Shedding Light on the Web of Documents. In *Proceedings of the 7th International Conference on Semantic Systems - I-Semantics '11*. ACM Press, Graz, Austria, 1–8. https://doi.org/10.1145/2063518.2063519

[26] Federico Nanni, Bhaskar Mitra, Matt Magnusson, and Laura Dietz. 2017. Benchmark for Complex Answer Retrieval. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. ACM, Amsterdam The Netherlands, 293–296. https://doi.org/10.1145/3121050.3121099

[27] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *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 (CEUR Workshop Proceedings, Vol. 1773)*, Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (Eds.). CEUR-WS.org.

[28] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-Ranking with BERT. *CoRR* abs/1901.04085 (2019). arXiv:1901.04085

[29] Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 803–818. https://doi.org/10.18653/v1/2020.findings-emnlp.71

[30] Marco Ponza, Diego Ceccarelli, Paolo Ferragina, Edgar Meij, and Sambhav Kothari. 2021. Contextualizing Trending Entities in News Stories. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. ACM, Virtual Event Israel, 346–354. https://doi.org/10.1145/3437963.3441765

[31] Chen Qu, Liu Yang, Minghui Qiu, W. Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019. BERT with History Answer Embedding for Conversational Question Answering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 1133–1136. https://doi.org/10.1145/3331184.3331341

[32] Chen Qu, Liu Yang, Minghui Qiu, Yongfeng Zhang, Cen Chen, W. Bruce Croft, and Mohit Iyyer. 2019. Attentive History Selection for Conversational Question Answering. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu (Eds.). ACM, 1391–1400. https://doi.org/10.1145/3357384.3357905

[33] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994 (NIST Special Publication, Vol. 500–225)*, Donna K. Harman (Ed.). National Institute of Standards and Technology (NIST), 109–126.

[34] Tetsuya Sakai. 2014. Statistical Reform in Information Retrieval? *ACM SIGIR Forum* 48, 1 (June 2014), 3–12. https://doi.org/10.1145/2641383.2641385

[35] Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do Neural Dialog Systems Use the Conversation History Effectively? An Empirical Study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 32–37. https://doi.org/10.18653/v1/P19-1004

[36] Diogo Tavares, Pedro Azevedo, David Semedo, Ricardo Sousa, and Joao Magalhaes. 2023. Task Conditioned BERT for Joint Intent Detection and Slot-filling. In *Progress in Artificial Intelligence - 22nd EPIA Conference on Artificial Intelligence, EPIA 2023, Faial, Portugal, September 5 - 8, 2023, Proceedings*. Springer.

[37] Svitlana Vakulenko, Nikos Voskarides, Zhucheng Tu, and Shayne Longpre. 2021. A Comparison of Question Rewriting Methods for Conversational Passage Retrieval. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12657)*, Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). Springer, 418–424. https://doi.org/10.1007/978-3-030-72240-1_43

[38] Christophe Van Gysel, Manos Tsagkias, Ernest Pusateri, and Ilya Oparin. 2020. Predicting Entity Popularity to Improve Spoken Entity Recognition by Virtual Assistants. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Virtual Event China, 1613–1616. https://doi.org/10.1145/3397271.3401298

[39] Johannes M. van Hulst, Faegheh Hasibi, Koen Dercksen, Krisztian Balog, and Arjen P. de Vries. 2020. REL: An Entity Linker Standing on the Shoulders of Giants. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Virtual Event China, 2197–2200. https://doi.org/10.1145/3397271.3401416

[40] Chenyan Xiong, Jamie Callan, and Tie-Yan Liu. 2016. Bag-of-Entities Representation for Ranking. In *Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval, ICTIR 2016, Newark, DE, USA, September 12-6, 2016*, Ben Carterette, Hui Fang, Mounia Lalmas, and Jian-Yun Nie (Eds.). ACM, 181–184. https://doi.org/10.1145/2970398.2970423

[41] Chenyan Xiong, Zhengzhong Liu, Jamie Callan, and Tie-Yan Liu. 2018. Towards Better Text Understanding and Retrieval through Kernel Entity Salience Modeling. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz (Eds.). ACM, 575–584. https://doi.org/10.1145/3209978.3209982

[42] Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2020. Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 23–30. https://doi.org/10.18653/v1/2020.emnlp-demos.4

[43] Jheng-Hong Yang, Sheng-Chieh Lin, Chuan-Ju Wang, Jimmy Lin, and Ming-Feng Tsai. 2019. Query and Answer Expansion from Conversation History. In *Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, November 13-15, 2019 (NIST Special Publication, Vol. 1250)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST).

[44] Liu Yang, Junjie Hu, Minghui Qiu, Chen Qu, Jianfeng Gao, W. Bruce Croft, Xiaodong Liu, Yelong Shen, and Jingjing Liu. 2019. A Hybrid Retrieval-Generation Neural Conversation Model. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu (Eds.). ACM, 1341–1350. https://doi.org/10.1145/3357384.3357881

[45] Chen Zhang, Hao Wang, Feijun Jiang, and Hongzhi Yin. 2021. Adapting to Context-Aware Knowledge in Natural Conversation for Multi-Turn Response Selection. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 1990–2001. https://doi.org/10.1145/3442381.3449902

[46] Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020. Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2031–2043. https://doi.org/10.18653/v1/2020.acl-main.184

[47] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 1441–1451. https://doi.org/10.18653/v1/p19-1139

[48] Chen Zhao, Chenyan Xiong, Corby Rosset, Xia Song, Paul N. Bennett, and Saurabh Tiwary. 2020. Transformer-XH: Multi-Evidence Reasoning with eXtra Hop Attention. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.