Information Sciences 607 (2022) 1–26

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Qualitative measures for ad hoc table retrieval

Maryam Khodabakhsh [a,*], Ebrahim Bagheri [b]

[a] Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran
[b] Laboratory for Systems, Software and Semantics (LS^3), Ryerson University, Canada

ARTICLE INFO

Article history:
Received 19 October 2021
Received in revised form 27 April 2022
Accepted 23 May 2022
Available online 31 May 2022

Keywords:
Qualitative measures
Ad hoc table retrieval
Deep contextual embeddings

ABSTRACT

The focus of our work is the ad hoc table retrieval task, which aims to rank a list of structured tabular objects in response to a user query. Given the importance of this task, various methods have already been proposed in the literature that focus on syntactic, semantic and neural representations of tables for determining table relevance. However, recent works have highlighted queries that are consistently difficult for baseline methods to satisfy, referred to as *hard queries*. For this reason, the objectives of this paper include: (1) effectively satisfying hard queries by proposing three classes of qualitative measures, namely coherence, interpretability and exactness, (2) offering a systematic approach to interpolate these three classes of measures with each other and with baseline table retrieval methods, and (3) performing extensive experiments using a range of baseline retrieval methods to show the feasibility of the proposed measures for hard queries. We demonstrate that the consideration of the proposed qualitative measures will lead to improved performance for hard queries on a range of state-of-the-art ad hoc table retrieval baselines. We further show that our proposed measures are synergistic and will lead to even higher performance improvements over the baselines when interpolated with each other. The improvements measure up to 22.94% on the Semantic Table Retrieval (STR) method with an NDCG@20 of 0.5, which is superior to the performance of any state-of-the-art baseline for hard queries in the ad hoc table retrieval task.

1. Introduction

Given their expressive form and ease of interpretability, tables are widely used to report on a range of topics like statistical data, financial reports, business information, and even cultural topics and are effectively used to satisfy user information needs [1,2]. Tables are so prevalent on the web that Cafarella et al. [3] were able to extract 14.1 billion HTML tables based on Google's common crawl. For this reason, web table mining has gained increasing attention from the research community in recent years [2,4–6] by covering tasks like table retrieval [3,6], table augmentation [7], web table embeddings [8], and query performance prediction [9]. The focus of our work in this paper is on ad hoc table retrieval, which is concerned with the identification and ranking of the most relevant tables for a given user query. As argued by Zhang et al. [6], the effective retrieval of tables can impact other table-mining tasks such as table completion and augmentation.

While ad hoc table retrieval is a subset of the ad hoc retrieval task, it has its own unique challenges that distinguish it from traditional ad hoc document retrieval tasks. For instance, unlike documents, tables lack context and continuity, which

* Corresponding author.
E-mail addresses: m_khodabakhsh@shahroodut.ac.ir (M. Khodabakhsh), bagheri@ryerson.ca (E. Bagheri).

https://doi.org/10.1016/j.ins.2022.05.080
0020-0255/

are common elements when determining relevance in ad hoc retrieval. Therefore, traditional methods such as language models [10] or frequency-based term-matching models such as BM25 [11], which view a table as a single field document, do not show strong retrieval effectiveness on this task [3]. The literature shows that learn to rank (LTR) models that operate based on hand-crafted features from query and table spaces are among the most effective retrieval methods [**?**,6]. While effective in terms of performance, they rely on manually defined features and hence are not necessarily scalable across a range of domains. For this reason, more recent work has focused on deep ranking methods that learn suitable representations for tables without the need for manual intervention [12–14].

Despite the increasing effectiveness of both LTR models and deep ranking methods, Bagheri et al. [4] have shown the effective performance of existing methods is due to their effectiveness on a subset of the query space. In other words, these methods do not show a uniform performance across a range of queries. There is a set of *hard queries* that is difficult to address and hence is often not addressed effectively by baseline methods. The **objective of this paper** is to develop techniques that improve the performance of existing methods over such *hard queries*.

In the context of ad hoc table retrieval, hard queries are often those that consist of terms not appearing in their relevant tables, which means the query and the relevant tables have low lexical overlap. This means that (1) traditional retrieval methods based on frequency term matching models, such as BM25, suffer from the lexical chasm problem and are unable to satisfy such hard queries, and (2) neural retrieval techniques also show a weak performance on such hard queries as they often only capture semantics of terms from pre-trained neural embeddings without contextualizing table structure and the association between different table modalities.

The main objective of our work in this paper is to improve the performance of such hard queries that are difficult to address by a range of state-of-the-art baseline methods. We hypothesize that there are quality characteristics for each table that makes them more desirable when compared with other tables for the users and hence impact the users' perception of the relevance of a table. We refer to such quality characteristics as qualitative measures and show that considering these qualitative measures leads to notable performance improvements over both all and hard queries. We have defined the qualitative measures in such a way that the structure, syntactics, semantics, and contextual information of tables are considered. These aspects are overlooked by existing frequency or neural methods.

To address such hard queries, while existing methods for table retrieval primarily focus on developing or learning a measure of relevance, we propose that using any table for a user query depends on factors beyond relevance. Some qualitative aspects of a table impact whether a user finds a table appropriate for their information needs. We propose to augment the performance of existing state-of-the-art ranking methods by incorporating a set of table-specific qualitative measures when ranking tables. We classify these qualitative measures into three classes, namely *coherency*, *interpretability*, and *exactness*. We propose that the effectiveness of ranking methods could be improved by considering tables' qualitative characteristics when ranking them in response to a user query. More specifically, we hypothesize that among the tables retrieved by a baseline ranking method, those that have certain desirable qualitative characteristics have are more likely to satisfy a user's information needs.

We formalize the three classes of qualitative characteristics into coherency, interpretability, and exactness, as follows:

- *Coherency* refers to the consistency of information provided within a given table and its different modalities. As such, we hypothesize that a table with a higher degree of coherency would be more desirable for a user and more likely to satisfy their information needs.
- *Interpretability* is the degree to which the information presented in a table is understandable. Given tables often present information in a structured yet compact form, a table that holds contextual information or one that provides access to contextual information would be easier to understand and hence likely to be more desirable to users.
- *Exactness* is how precisely a table can correspond to a user's information needs. Most tables provide information or data on a range of topics. While these topics are often related to each other, they do not all necessarily satisfy the information needed by the user. A table with the exact information related to a user's query would allow them to find the information they are looking for.

We offer concrete formalization of how to measure the three qualitative characteristics from different perspectives and employ them to re-rank the retrieved list of tables by state-of-the-art, ad hoc table retrieval methods and show that re-ranked tables have significantly higher retrieval effectiveness than hard queries that are challenging for the baseline methods to address. Our evaluations are performed based on the WikiTables test collection introduced in [6].

In summary, our work makes the following key contributions:

1. Systematically introduce, classify, and formalize sets of the novel qualitative measures of a table's characteristics to estimate the likelihood of a table satisfying a user's information needs.
2. Show how each of these qualitative measures can be integrated into a re-ranking process to obtain a more effective table ranking.
3. Evaluate our proposed qualitative measures on a gold standard test collection derived from Wikipedia tables. We further analyze our findings from various perspectives and identify areas where qualitative measures can provide significant performance improvements.

In the next section, we concisely discuss and review the related work in the context of ad hoc table retrieval. Sections 3 and 4 provide the technical details of our proposed qualitative measures. The details of our experiments, gold standard test collection, baselines, and our findings are presented in Section 5. Section 6 is dedicated to a discussion of our research findings. Finally, the paper is concluded with Section 7.

2. Related work

The focus of this paper is the task of searching, retrieving, and ranking tables in response to an unstructured user query, often known as ad hoc table retrieval. Earlier table retrieval methods such as TableRank [15] and WebTables [3] viewed tables as a single field document and applied traditional document retrieval methods [10,11]. Subsequent work adopted a similar strategy but shifted towards multimodal retrieval where different structural elements of each table were viewed as a different modality, e.g., title, caption, headers, rows, and columns [16,17,5]. These approaches, while the first to attempt an ad hoc table retrieval, did not show satisfactory performance on the task.

Further methods focused on manually crafted features extracted from the query and table spaces that could be used in a learning to rank framework. These features can be classified as *table-specific*, *query-specific*, and *table-query interdependency features* [3,7]. Table-specific features include those that capture the specific properties of a table, like the number of empty table cells or the number of times the table has been viewed in the past. Query-specific features represent a query's characteristics like its length. The third class of features, namely table-query interdependency features, measure the association between a query and table pairs, e.g., the frequency of query terms in the table body. More recent techniques capture the semantic aspects of tables and queries in the table ranking process [**?**,6,18]. Zhang and Balog [6,18] were the first to develop a ranking method to match queries and tables based on their semantic association where both queries and tables were represented using semantic concepts (bag-of-entities and bag-of-categories) as well as continuous dense vectors (word and graph embeddings). Additional work was subsequently developed to learn neural representations for tables for the purpose of retrieval [5]. The semantic and neural features are also used in the context of learning to rank.

To avoid the need to define hand-crafted features, Shraga et al. [14] and Li et al. [19] developed a multimodal deep-learning approach for web table and image retrieval, respectively, as cross-modal retrieval [20]. Li et al. [19] propose a novel weakly-supervised deep embedding model using a matrix factorization framework for collaborative social image understanding, called Deep Collaborative Embedding (DCE), which can simultaneously address multiple retrieval tasks like image-to-tag retrieval (tag refinement and assignment), CBIR, TBIR, and tag-to-tag retrieval (tag expansion). The proposed DCE method effectively estimates the relevance between tags and images. Shraga et al.'s approach [14] assumes that each structural element of a table is a unique modality that can play a different role in determining table relevance to a query. Based on this multimodal approach, the authors apply a gating scheme to learn a joint multimodal table representation that is used to measure table and query associations. In context of table retrieval, Shraga et al. [14] developed a multimodal deep-learning approach for web table retrieval. This approach assumes that each structural element of a table is a unique modality, which can play a different role in determining table relevance to a query. Based on this multimodal approach, the authors apply a gating scheme to learn a joint multimodal table representation that is used to measure table and query associations.

Unlike the work by Shraga et al. [14], Chen et al. [13] piggyback on pre-trained BERT representations to construct a representation for each table, however, this approach is computationally expensive. To reduce computation time, Trabelsi et al. [12] proposed the DSRMM model to include summary vectors about the contents of the table, both in terms of values in each column and values in select rows. The summary vectors compress each row and column into a fixed-length feature vector using word embeddings of data values. There are two advantages to using summary vectors, first, it reduces computation time, second, computing the summary vectors is independent of the query, therefore, table representations can be learned offline. Recently, a Graph-based Table Retrieval (GTR) [21] framework was introduced in which a table is first converted into a tabular graph, with cell nodes, row nodes, and column nodes to capture content at different granularities. Then, the tabular graph is input to a graph transformer model that can capture both table cell content and the layout structures.

The two-step retrieval process [22] is gaining popularity in the ad hoc retrieval community where first, a traditional IR method, such BM25 [11] or query likelihood [10], retrieves top-k documents from a given collection. Then, a computationally expensive model like a neural ranking model, is used to re-rank the top k documents from the initial retrieval step. During the first step, recall is more important than precision to cover all possible relevant documents. In the second step, the high recall retrieved list is re-ranked to optimize for ranking precision. Shraga et al. proposed TabIESim [16] for re-ranking the (tables) ranked lists which are obtained by state-of-the-art table retrieval methods. They consider a new combination of intrinsic and extrinsic (similarity) sources using passage-level information and a regularized manifold-based ranking approach, respectively. Also, they introduced another table re-ranking model named PTRM [17] that builds on top of the relevance model. They utilized table columns as pseudo-relevance feedback to blindly expand a given user's query.

Despite the notable performance improvements shown by recent ad hoc table retrieval methods, Bagheri et al. [4] have shown that a set of queries, often known as *hard queries*, is not effectively addressed by existing retrieval methods. In other words, while performance improvements are observed by baseline methods, they are not uniformly observable across all queries. We note that this issue is not unique to ad hoc table retrieval as researchers have already reported and extensively studied the issue of hard queries over a range of retrieval domains including ad hoc document retrieval [23], question answering [24], Coreferent Mention Retrieval (CMR) [25], and entity-based retrieval [26]. Given the significance of effectively

addressing hard queries, various earlier works attempted to predict the performance of each specific query before being addressed. This includes our work on query performance prediction for ad hoc table retrieval [9] and other more recent work on neural-based methods for query performance prediction in other retrieval domains [27–29]. In one very recent work on hard queries, Arabzadeh et al. [30] argue that not all queries can be addressed effectively through different retrieval methods and hence propose to adopt different retrieval strategies at runtime depending on each query.

Our work in this paper extends earlier work on hard queries in the broader area of hard queries in information retrieval and specifically focuses on hard queries in ad hoc table retrieval. More specifically, the objective of this paper is to propose a re-ranking framework to re-score candidate tables based on three classes of qualitative measures, which capture the specific characteristics of relevant tables that would not be otherwise captured when addressing hard queries. We will show that by using such qualitative measures, together with state-of-the-art table retrieval methods, the performance of hard queries will significantly improve over existing state-of-the-art baselines.

3. Proposed framework

For our proposed table re-ranking framework we first formally define the re-ranking framework and describe tables as multimodal objects.

3.1. Table re-ranking

Let q, T, and $R(q,T)$ denote a query, a table in table corpus τ, and T's relevance score to q assigned by a base ranking function, respectively. The ranking function returns a ranked list of top k tables $(T_1, T_2, \ldots, T_k)$ from τ satisfying: $R(q,T_i) >= R(q,T_{i+1}), \forall i \in \{1,2,\ldots,k-1\}$.

Within our re-ranking framework, given an initial pool of k table candidates retrieved by a base ranking function (R), the tables are re-scored through the interpolation of the relevance score from the base ranking function and some qualitative measure of table characteristics that will be introduced later in this paper. We obtain the final table score through a linear interpolation approach:

$$Score(q,T_i) = \alpha R(q,T_i) + \beta f^{QM}(q,T_i) \qquad \forall i \in \{1,2,\ldots,k-1\} \tag{1}$$

where α and $\beta = 1-\alpha$ are linear interpolation coefficients so that $\alpha + \beta = 1$. $R(q,T_i)$ is the normalized value (through min–max feature scaling) of the baseline ranking function that provides the relevance score of T_i for q. $f^{QM}(q,T_i)$ is a function that computes some qualitative measures (QM) for table T_i. Based on $Score(q,T_i)$, we define our re-ranking framework as $Score(q,T_i) >= Score(q,T_{i+1})$.

3.2. Tables as multimodal objects

We view each table $T \in \tau$ as a multimodal object represented as five tuples $T = \{description, schema, row, column, cell\}$ where $description, schema, row, column$, and $cell$ are the values obtainable from the corresponding elements of the table. More specifically, each of these tuples consists of the following information:

1. *Description* constitutes information like the table's caption and any titles of the document or web page where the table appeared. Each table's description is commonly written in natural language textual form.
2. *Schema* is a set of N column heading labels $\{l_1, l_2, \ldots, l_N\}$. Unlike a table's description, the information in the schema text is usally short and may be abbreviated.
3. *Row* is a set of M elements denoted as $row = \{r_1, r_2, \ldots, r_M\}$ where r_i is defined as the concatenated form of the cells observed in row i appended from left to right.
4. *Column*, like Row, is a set of N elements $column = \{c_1, c_2, \ldots, c_N\}$, which groups the table data vertically and concatenates data in each column from top to bottom.
5. *Cell* consists of unrolled matrix of size $M * N$ cells, denoted as $cell = \{cell_1, cell_2, \ldots, cell_{M*N}\}$.

We view each of these tuples as a modality of the table.

4. Proposed qualitative measures

In this paper, we introduce three main classes of qualitative measures: coherency, interpretability, and exactness. These three classes are proposed based on the hypothesis that the qualitative characteristics of a table influence how users decide on the utility and relevance of the table for their information needs. For instance, we hypothesize that those tables that consist of more coherent content are more desirable tables for the users and therefore have a higher likelihood of being relevant. We introduce and formalize each of these three classes of qualitative measure in the below subsections.

4.1. Table coherency

Coherency measures how well the content within one or more modalities of a table relate to one another. Intuitively, a coherent table is one where the modalities are all tightly related to each other and the content within each modality is consistent. Let us illustrate the concept of coherency through a sample query 'Broadway musicals director'. We illustrate two tables that could be related to this query in Figs. 1 and 2 and. Based on the gold standard relevance judgment dataset (introduced in the experiments section), the table in Fig. 1 is considered to be the true relevant table for this query, while the table in Fig. 2 is not considered relevant. We find that while both tables include the title of the play, the irrelevant table consists of information about movies that were later developed based on the given play. However, the relevant table is focused solely on the play and provides information about the year, title, and director. As such, the table in Fig. 1 consists of more coherent information and is hence more desirable for users.

The notion of *coherency* has already been investigated by other researchers in other contexts. For example, Zheng et al. [31] introduced a cross-modal retrieval method based on the coherency between the semantic structure of multiple modalities. The modality gap is the main challenge of cross-modal retrieval and a common approach to bridging the modality gap is by constructing a shared representation space where the multimodal samples can be represented uniformly. The authors argue that it is not easy to address the modality gap because it requires detailed knowledge of the content of each modality and the correspondence between them, which they address through *coherency*. Tu et al. [32] also propose a set of coherency features to improve information retrieval performance. In their approach, each document is represented as a graph-of-words that corresponds to a weighted directed graph whose vertices and edges represent unique terms and the degree of semantic relatedness between the term, and the other terms within the document, respectively. Coherency is measured as a function of this graphical representation. Similarly, Wang et al. [33] have investigated proposing a coherency-based retrieval approach. The authors designed a symmetrical convolutional neural network to capture semantic coherence within a learning to rank framework. Our work in this paper is inspired by earlier works on *coherency* in retrieval and hypothesizes that when presented with competing tables such as those shown in the 'Broadway musicals director' query example, a more effective retrieval method would need to prioritize more coherent tables over their counterparts.

Given table T and a special *modality* of T, the coherency measure of T is defined as follows:

$$f^{Coherency}(q,T) = Aggr_{a \neq b \in modality} sim(a,b) \qquad \forall modality \in T \tag{2}$$

where $sim()$ is a function that computes the similarity between two textual items a and b. We formally define coherency from three different perspectives introduced in the following:

4.1.1. Statistical coherency

The simplest form of coherence can be defined based on lexical association of content within different modalities. If the content in different modalities has similar lexical representations, then it is likely that the content in the whole table is coherent. To this end, we adopt a lexical matching function, e.g., BM25, as $sim()$ to measure the consistency of the content of the different modalities of a table. In our work, we adopt the symmetric variant of BM25 for $sim()$. More specifically, for the statistical coherency of T, the $Aggr()$ function is defined as the average pairwise lexical association between content in different modalities.

4.1.2. Semantic coherency

While statistical coherency determines the consistency of table content based on their lexical association, relying on pure lexical association can overlook the semantic relationship between table content for cases when the same information is pre-

Year	Title	Director(s)
2009	"Same Thing We Do Everyday Pinky" (featuring Craig Owens)	Robby Starbuck
2011	"Last Saturday"	Levon Mergian

Fig. 1. A snapshot of the relevant table for the 'Broadway musicals director' query.

Play	Playwright	Film	Film director
8 femmes (1958)	Robert Thomas	*8 Women* (2002)	François Ozon
The 24th Day	Tony Piccirillo	*The 24th Day* (2004)	Tony Piccirillo
27 Wagons Full of Cotton (1946), and *The Longest Stay Cut Short* or *The Unsatisfactory Supper* (1946)[3]	Tennessee Williams	*Baby Doll* (1956)	Elia Kazan[4]

Fig. 2. A snapshot of a portion of a related table that is not considered relevant to the 'Broadway musicals director' query by the human judges.

sented with different terminology (e.g., vocabulary mismatch) [34]. To operationalize sim(), we introduce two strategies, based on 1) neural embeddings, and 2) deep contextual embeddings.

Neural embeddings: Researchers such as Zamani et al. [35] and Bagheri et al. [23] have shown that it is possible to use pre-trained neural embeddings to improve the task of ad hoc document retrieval. Such works have argued that while not always effective, the interpolation of pre-trained neural embeddings with lexical matching techniques often results in more effective retrieval. On this basis, we propose to define $sim()$ as a function that determines the semantic association between the content of the table modalities based on the vector similarity of the neural embedding representations. There are two main approaches for forming embedding representations for free form textual content based on neural embeddings, namely Paragraph Vectors (PV) and Bag-of-Word-Embeddings (BoWE).

In the first approach, PV, the concept of paragraph vectors [36] has been introduced for learning high-quality distributed representations for textual content. Le and Mikolov [36] have shown that PVs capture document semantics through dense vectors. This is particularly useful for our case, where vector representation of modalities can be developed based on their paragraph vector representation, i.e., $sim(a,b) = Cos(PV(a), PV(b))$. The second approach adopts a BoWE [37] strategy where each document (textual snippet) is represented as a set of unordered embedding representations of its constituent terms. Based on this representation, it is possible to formally define $sim(a,b) = \max_{w_a \in a, w_b \in b}\left(\left\{cos\left(\overrightarrow{w_a}, \overrightarrow{w_b}\right)\right\}\right)$ where w_a and w_b are terms in a and b. Like statistical coherency, the $Aggr()$ function is defined as the average pairwise semantic association between the textual content.

Deep contextual embeddings: Unlike neural embeddings in the previous section, such as paragraph vectors, contextual embeddings allow us to extract sequence-level semantics by using context to determine the best representation. This is especially meaningful in the context of table coherency as we are interested in knowing whether the contents are consistent in the context in which they appear. To formalize semantic coherency based on deep contextual embeddings, we hypothesize that a coherent table is one that presents content in each of its modalities that are consistent with the content contained in the table's description, e.g., table caption. Simply put, the content in each of the table modalities should be interpreted in the context of the table description.

For this purpose, we adopt the Next-Sentence Prediction (NSP) task in transformer-based language models such as BERT [38] to make judgments about the relationship between the table description and the different table modalities. The NSP task takes two sentence embeddings (sentence A and sentence B) and predicts whether sentence A immediately precedes sentence B by encoding two meta-tokens ([SEP] and [CLS]). The [SEP] token separates the tokens of two sentences, and the [CLS] token is used for making judgments about the text pairs. The [CLS] token's representation is fed into the softmax layer to get the probability that B follows A. We encode the table description and each of the table modalities as sentences A and B, respectively.

In the semantic coherency of T based on deep contextual embeddings, a is T's description and b can be any of the table modalities, except its description, as that is used as contextual information. Here, NSP as $sim()$ denotes the likelihood that the second sentence (content from table modality) follows the first (content from table description). The Aggr() function integrates over multiple measurements comparing sentences from table description to sentences from the modality.

One could argue that semantic coherency is a generalization of statistical coherency that allows for a more flexible account of association between modalities by allowing modality content to be related to each other, even if they do not have the exact same lexical representation.

4.1.3. Topical coherency

The idea behind topical coherency is to assess the consistency of content within a table based on the topic distributions observed in the content in table modalities. Unlike statistical coherency, which measures the lexical association between modalities, and semantic coherency, which computes the soft similarity between modality content, topical coherency focuses on determining whether the content in table modalities belong to similar topics or not. A topically coherent table would be one with contents belonging to the same or similar topics derived through a topic model such as the Latent Dirichlet Allocation.

To formalize topical coherency, we adopt a similar assumption to that of the deep contextual semantic coherence measure where we consider a table to be coherent if the topic distributions observed in the table content are consistent with the topic distribution observed in the table description. To this end, given the topic distributions observed in both the table description and table contents are in the form of probability distributions, we employ the symmetric variant of Kullback–Leibler (KL) divergence to quantify the similarity between these topic distributions [39] as $sim()$. The description and content of the table are a and b, respectively.

A topically coherent table would be one in which the topic distributions observed in the table description align with those observed in the table itself. As we will explain later, topic distributions are learned over the whole table corpus using the latent dirichlet allocation method. Topic distributions in both the table description and the table itself is inferred using Gibbs sampling.

4.2. Table interpretability

As mentioned in the literature [39], a distinguishing aspect of table retrieval compared with ad hoc document retrieval is that tables often lack context where information presented in a table are assumed to carry their own semantics through

either the position where they appear in relation to (1) table schematic information or (2) other information around the table or linked by the table through hyperlinks. In the first class of table characteristics, i.e., table coherency, we have already captured the table context as represented through their schematic information. This is accomplished by determining whether item of information in different table modalities are consistent with each other. Consistency across modalities shows that different modalities provide contextual information to support each other and hence the table has a higher likelihood of being interpretable for users. Now, in the context of table *interpretability*, we are interested in proposing measures that indicate how much contextual information is available for the interpretation of the table through its surrounding content.

We hypothesize that a table that provides a higher degree of context will offer the user wider access to additional information about the table content and, thus, leads to better interpretability resulting in a higher degree of utility for the user. In principle, one can view table interpretability as a form of table relevance where the relevance of the table to the query is computed based not on the content of the table but rather on the combination of the table content and its extension. As such, a truly relevant table that might not be considered relevant in itself, would be deemed to be relevant once additional contextual information was added to it. Similarly, a table considered to be relevant could be identified to be less relevant once its associated content is added to it. As such, interpretability can help measure table and query relevance by extending each table's contextual information.

Researchers have already suggested enriching incomplete textual data representations [40–43] by exploiting external resources such as WordNet [44], Wikipedia [44,45], and data from the Open Directory Project (ODP) [46] to increase understandability and interpretability during retrieval. For instance, Li et al. [41] proposed using Wikipedia as an external knowledge source to identify top-k most relevant Wikipedia articles to be used as contextual information. Pan et al. [47] explored enhancing Question Answering (QA) methods with contextual information from background knowledge sources. For this purpose, they investigated the impact of pre-trained language models for identifying relevant concepts that could be added to QA methods as background knowledge. A more recent method, known as MAVEx [48], used both textual and visual external knowledge resources, including images searched using Google, sentences from Wikipedia articles, and concepts from ConceptNet to add additional contextual information to QA systems to improve QA effectiveness. Similarly, our work in this paper focuses on providing additional contextual information when performing table retrieval through the *interpretability* qualitative measure.

To define table interpretability, we propose capturing the amount of context available for a table by considering how contextual information can be extracted for the table. We extend each table based on its contextual information through two table extension strategies, namely 1) the *surrounding text-based* method, which considers the content that appears on the same page as the table to be the extension of the table, and 2) the *link-based* method that employs the content of pages linked to the table as extensions of the table. Let us visually present a sample query and its relevant table to show that it would be very difficult to retrieve this relevant table for the query unless an interpretability measure were employed. We consider the 'Apples market share' as the sample query and show its relevant table in Fig. 4. As shown in the figure, the relevant table does not provide much information to be effectively matched against the query. However, the text in the first column consists of links to external sources, e.g., the link in the first cell of the first row points to an IDC report on 'Smartphone Market Share'. This reports discusses various smartphone manufacturer market shares at length. Therefore, if such information is used to extend table context, it would be more likely that this table would be considered relevant for the query. In such cases, interpretability measures would enable us to retrieve relevant tables more effectively.

The two *surrounding text-based* and *link-based* extension strategies allow us to extend table T to an extended form, $ext(T)$. Given $ext(T)$, we develop a generic framework for determining the impact of table extension on table interpretability and hence retrieval effectiveness. We measure the interpretability of Table T in the context of query q as: Fig. 3

$$f^{Interpretability}(q,T) = sim(q, ext(T)) \tag{3}$$

A table will be more interpretable for a user looking for information about q when the extension of the table has closer resemblance to q.

To operationalize *table interpretability*, we benefit from similar techniques used for table coherency for computing the $sim()$ function in Eq. 2:

- The first strategy would be to adopt a lexical matching strategy, which we refer to *statistical interpretability*. This is like statistical coherency in that it employs the lexical association between the query and the table extension to determine the degree of table interpretability after being extended. In contrast to table coherency that computes the lexical association between different table modalities, table interpretability compares the extension of the table to the query.
- The second strategy would be to benefit from the semantic association between the query and the table extension. Akin to semantic coherency, we adopt both neural embeddings and deep contextual embeddings to determine *semantic interpretability*. Like neural embeddings in semantic coherency, we consider both BoWE and PV representations to extract representations for the query and the extension of the table. These representations can then be used to measure the association between the user query and the extension of the table. We also adopt deep contextual embeddings to learn representations for both the input query and the table extension. Like deep contextual coherence, which measures whether the content of one modality is an expected continuation of another modality, here we employ deep contextual embeddings and the NSP task to determine whether the extension of the table is a likely continuation of the input query.

IDC: Worldwide smartphone shipments (millions of units)

Quarter	Android	iOS	Windows	BlackBerry	Symbian	Other	Total
2017 Q1[214]	292.7	50.6	0.34	–	–	0.34	344.3
2016 Q4[214]	318.3	71.2	0.78	–	–	0.78	391.0
2016 Q3[215]	315.3	45.4	0.9	–	–	1.6	363.2
2016 Q2[215]	302.7	40.4	1.4	–	–	1.0	345.5
2016 Q1[215]	291.3	53.8	2.79	–	–	1.40	349.3
2015 Q4[215]	291.7	68.5	4.40	–	–	1.83	366.4
2015 Q3[216]	329.04	46.70	14.67	–	–	3.94	394.35

Fig. 3. A snapshot of a portion of the relevant table for the '*apples market share*' query.

Game name	Year	Origin	Players	Gameplay style	Similar Games	Reference
Love Letter	2012	Kanai Factory	2–4	Risk and deduction game	Coup	
Gomoku (五目並べ, *gomokunarabe*)	circa 850	Traditional	2	Strategic abstract game played with Go pieces on a Renju board (15×15), goal to reach five in a row	Renju, Four in a row	
Jinsei Game (人生ゲーム, *jin-sei gēmu*)	1967	Takara	?	Japanese adaption of The Game of Life	The Game of Life	
Machi Koro (街コロ)	2012	Grounding Inc.	2–5	Tabletop city-building/resource-gathering game using cards and dice	Catan	
Renju (連珠)	1899	Traditional	2	Strategic five-in-a-row game with equal chances for both players	Pente, Gomoku	

Fig. 4. A snapshot of a portion of the relevant table for the '*board games number of players*' query.

These two table extension strategies allow us to measure the semantic association between a user's query and the extension of the table. We hypothesize that a higher association between the query and the extended table indicates higher interpretability of the table through its context and hence such a table would have a better chance of being more desirable for users.

We note that unlike topical coherency, we do not employ topics to measure table interpretability, primarily because topics are effectively derived from longer documents and are not so effective for determining topics within search queries, which are often less than three terms [49]. Therefore, while it is possible to identify topics for table modalities and use them to measure their coherence, it will not be possible to effectively extract query topics to be used in comparison with the topics of the extension of the table.

4.3. Table exactness

Tabular structures often include information about many different aspects that are not all relevant to a given input query. For instance, a table that includes countries' GDPs is a relevant table to be retrieved for the 'Germany GDP' query, since it includes this information, however, there are other data in the table such as GDPs of other countries not directly related to the query. In such cases, the presence of other tangentially relevant information in the table could possibly lead to a lower desirability for the whole table and hence, as a result a lower retrieval performance. We hypothesize that it is possible to consider the presence of relevant information in table modalities as a sign of table relevance. In other words, a table can be considered to be relevant to a query to the extent of the relevance of its best matching modality.

On this basis, we propose the class of *table exactness* measures that compute the extent to which each table modality relates directly to the input query. Table exactness can be seen as a measure of *informativeness*, which shows how much specific information in a table one can maximally find if that table is retrieved for the query. One can view table exactness as a special case of table relevance for cases for which the most relevant segments of the table are identified and used to compute the degree of relevance, i.e., we pinpoint the table modality that is most related to the query and use that as a way to measure the association between the query and the table.

Other researchers have identified the need for capturing exactness in the related literature. For instance, Chen et al. [13] used the deep contextualized language model BERT to encode table content for the task of ad hoc table retrieval. In this context, the authors select a set of potentially informative modalities from the table as a more exact table representation. Zhou et al. [50] have also discussed the idea of extracting the most exact parts of a document for enhancing retrieval effectiveness.

They proposed a joint sentence scoring and selection framework to extract important sentences. Our work in the exactness measure aligns very well with the related literature and suggests that an effective retrieval would need to focus on the most concise parts of the table that relate to the query.

Let us consider a sample query and its relevant table to show how exactness can be helpful for more effective retrieval. Fig. 4 shows the relevant table for the 'board games number of players' query. When considering this table, one finds there are many more types of information about board games in this table beyond the number of players. The number of players of the board games is only one of the columns out of seven columns in the table. Therefore, in such a case, considering the information in the other columns would, in effect, reduce the likelihood of this table being relevant to the query. The proposed exactness measure will focus on the specific modality (in this case, column 4) that will maximize the retrievability of relevant tables for that query.

Given table T and a special *modality* of T, we formally define table exactness of T as follows:

$$f^{Exactness}(q,T) = max_{a \in modality} sim(q,a) \qquad \forall modality \in T \tag{4}$$

We develop $sim()$ from two different perspectives based on similar strategies to those adopted in the table coherency and table interpretability measures:

- When seeking information from tables, users are often looking for very specific data and/or information. Such data are usually embedded in a specific modality of the table. Therefore, while the relevance of the whole table to the query is important, the maximal relevance of table modalities could indicate how informative the table would be for the user. For this purpose, we propose the *statistical exactness* measure, which adopts a lexical matching strategy in which we convert the data in each table modality into a bag of words representation. We hypothesize that the most exact table is the one that consists of a modality that maximally shares the largest number of lexical overlaps with the query. In statistical exactness, $sim()$ is measure such as cosine similarity, q and each item of a modality are represented as n-dimensional vectors where n is the number of words in the vocabulary of the table corpus.
- Various researchers have already shown that soft matching strategies [23,35] could help connect queries to relevant content that may have been expressed in different terminology or presentation forms. Like table coherency and interpretability measures, we propose to measure the semantic exactness of a table given an input query based on the maximal soft matching between the query and the table modalities. Like coherency and interpretability measures, we perform soft matching, as required by $sim(q,a)$, based on the two strategies of (1) neural embeddings, and (2) deep contextual embeddings. In the neural embeddings strategy, we adopt the BoWE and PV. We also adopt the deep contextual embeddings strategy to perform soft matching based on a contextualized language model such as BERT [38]. Similarly, we adopt NSP to distinguish whether an item in the specific modality has the likelihood of following a given query.

Unlike the table coherency measure and like the table interpretability measure, we do not adopt topical information to measure exactness as topics are high-level representations of information domains and hence lack the specificity required to match queries to modalities at a finer-grained level. While appropriate for coherence to see if the information in different table modalities is related to each other, they are not as effective for short text such as queries [49].

Table 1 provides an overview of the measures that have been proposed for exploring different table characteristics. As described, the measures can be horizontally classified into measures coherency, interpretability, and exactness measures. In addition, each of these classes of measures are formalized based on lexical, semantic, and topical representations. Overall, this classification offers ten different quantifiable metrics for the three classes of qualitative table characteristics.

Table 1
Overview of the proposed qualitative table measures.

		Coherency	Interpretability	Exactness
Semantics	Statistical	Degree of lexical association between context in different table modalities.	The lexical similarity between the query and an extended form of the table.	The lexical correspondence between the query and the most similar item in the table modality.
	Neural embeddings	Degree of soft matching between different table modalities.	The degree of soft association between the query and an extension of the table.	The highest degree of soft matching between the query and the items in the table modalities.
	Deep contextual embeddings	Likelihood of table modalities being generated given the table description.	The likelihood of an extended table being generated given the input query.	The highest likelihood of one item in the table modality being generated given the input query.
	Topical	Consistency of topical distributions over the table content and its description.	*n/a due to the ineffectiveness of topic inference over short queries.*	*n/a due to the coarse-grained nature of topics.*

5. Experiments

The objective of our experiments is to evaluate the effectiveness of our proposed qualitative measures for improving the performance of hard queries in the context of ad hoc table retrieval. To this end, we define three Research Questions (RQs) that will be answered through our empirical experiments. Specifically, we address the following research questions in our experiments:

- *RQ1*. Which class of the proposed qualitative measures and what specific formalization of these measures would be most effective for re-ranking hard queries?
- *RQ2*. Would the most effective proposed qualitative measures lead to improved performance when interpolated with state-of-the-art methods in the table retrieval task in the context of hard queries?
- *RQ3*. Do the proposed qualitative measures have a synergistic impact on each other for table retrieval?

Below, we first introduce the test collection and the experimental setup and then report and analyze our results and present the most important findings in the context of our research questions.

5.1. Test collection

We employed the test collection introduced in [6] in our experiments. This test collection includes three main components, namely the table corpus, the query set and the relevance judgments. The table corpus is composed of the WikiTables corpus containing over 1.6 M tables. For each table, five information fields are provided: table caption, column headings, data rows, page title, and section title. The tables in the corpus have links to other Wikipedia pages that were extracted by Zhang et al. [6]. The query set consists of sixty queries, which were sampled from two independent sources. The relevance judgments consist of 3,120 query-table pairs. Out of these, 377 are labeled as highly relevant, 474 as relevant, and 2,269 as non-relevant. Akin to suggestions by earlier work [4], hard queries were considered to be those at the bottom 30% of queries ranked based on Normalized Discounted Cumulative Gain (NDCG) at a cut-off point of 20 for each baseline.

5.2. Experimental setup

In our experiments, we employ the pre-trained BERT [38] from HuggingFace with twelve layers, 768 hidden layers, twelve heads, and 110 M parameters trained on cased English text, often known as BERT-Base Cased. Furthermore, to implement BoWE and PV, we used GloVe[1] with 300-dimensions and doc2vec-wikipedia[2], both pre-trained on the full collection of English Wikipedia. We trained an LDA topic model on the latest dump of Wikipedia[3] with a topic size of 500. We note that when training was necessary, the results are reported based on fivefold cross validation (this was only necessary for the results in Table 14).

5.3. Ranking baselines

To select the state-of-the-art ranking baselines required in Eq. 1, i.e., $R(q, T)$, we adopt the most recent baselines that have shown the highest retrieval performance on the WikiTables corpus. These baselines include BRM [13], DSRMM [12], STR [6] and LTR [6]:

1. BRM has been recently proposed by Chen et al. [13] as a table retrieval method with three components. The first component is a content selector that extracts informative items (rows, columns, or cells) from a table. The second component uses BERT to extract features from the query, corresponding table context fields, and selected items. In other words, it constructs input for BERT considering the structure of the table. The output of BERT is fed into a regression layer as the last component to predict the relevance score. Chen et al. have three strategies to calculate the scores of items and three ways to construct items (as a list of columns, rows, or cells) from a table. Among all these variations, BERT-RowMax (BRM) achieves the best results across all metrics compared with the other variations.
2. DSRMM [12] is a recent method that combines deep contextual features with features based on term similarity distributions. The DSRMM method learns convolutional filters that extract contextual features from query/table interactions (semantic matching). This is combined with a feature vector based on the distributions of term similarity between queries and tables (relevance matching). Additionally, DSRMM incorporates table values using row and column summaries. Finally, it learns the contribution of each query token to the final relevance score. DSRMM is trained using a learning-to-rank approach with a listwise loss function.

[1] https://nlp.stanford.edu/projects/glove/
[2] https://github.com/jhlau/doc2vec
[3] https://dumps.wikimedia.org/enwiki/

3. STR [6] is a ranking function that models both the table and the query as sets of semantic vectors and then uses two general strategies (early and late fusion) for computing the similarity between queries and tables based on their semantic representations.
4. LTR [6] uses the full set of features listed in [7,3] and trains a Random Forest regression with 1,000 trees as the learner.

We note that we report all results based on NDCG@20 primarily because all baselines used in this paper report their results based on this metric, including BRM, DSRMM, STR, and LTR methods.

5.4. Visualization of hard queries

To demonstrate there is always a group of queries that is difficult for each table retrieval method, we visualize the performance of all the four baseline methods on the queries of the test collection in Fig. 5. In this figure, all queries are sorted based on their NDCG@20 performance in descending order such that queries with better performance are placed on the left of the figure while harder queries are placed on the right. We show that there is a group of queries in each method that has a performance of zero (on the right-most side of each figure) and a large number of queries with a performance lower than 0.5 (which we show using red brackets). As mentioned before, the objective of this paper is to systematically improve table retrieval over all baselines on such queries.

5.5. Findings

We present the findings of our experiments based on the three research questions introduced earlier.

5.5.1. RQ1. The effectiveness of the qualitative measures

The objective of the first research question is to explore which class of the proposed qualitative measures are impactful for improving the performance of hard queries in ad hoc table retrieval. Furthermore, we are interested in determining which formalization in each class of measures exhibit the best performance. To this end, we report the linear interpolation of the baseline methods with the proposed qualitative measures based on Eq. 1 below.

We first explore the impact of the proposed table coherence measure. We propose in this paper that table coherence can be formalized based on statistical coherence, semantic coherence, and topical coherence. Semantic coherence can be further operationalized by using neural embeddings and deep contextualized embeddings. Given earlier works [16] have shown synergistic impact between lexical and semantic scores for ad hoc retrieval, we interpolate statistical coherence and semantic and topical coherence measures. However, we note that for more comprehensive reporting of the results, the performance of each individual feature is reported separately in Appendix A.

The degree of impact of each of the table coherence measures are reported in Tables 2–4 for the different variations of table coherence measures. Based on Table 2, our first observation is that table modality plays an important role in how table coherency is measured and the final performance of the re-ranking process. As seen in the table, depending on which modal-

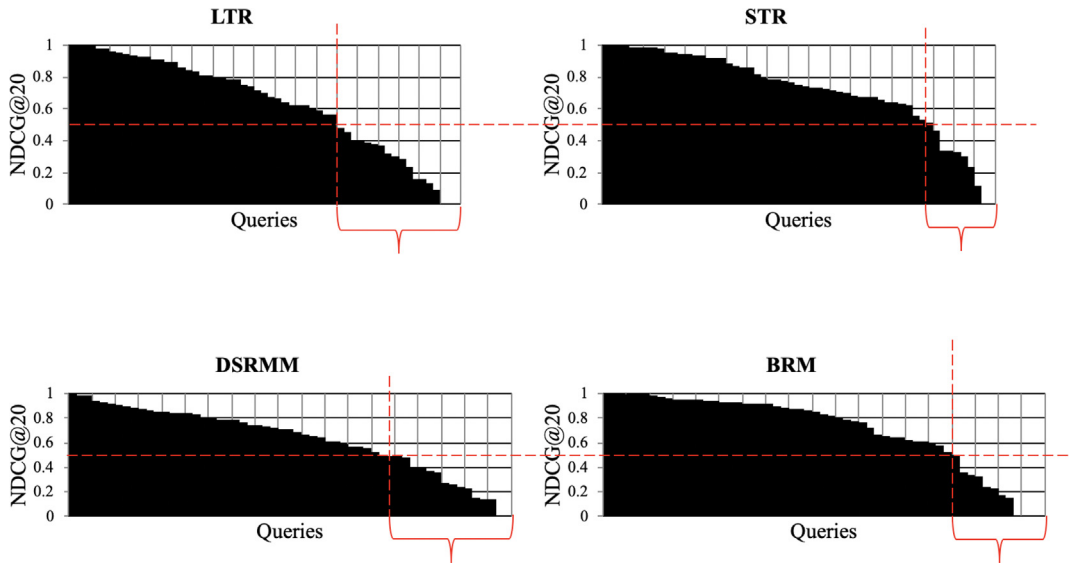

Fig. 5. Visualization of query performance over all queries. Queries are ranked based on their NDCG@20. We have denoted those sets of queries that have a performance of less than 0.5 on this metric using the red brackets.

Table 2
Percentage improvement by neural embedding coherence measures over hard queries in terms of *NDCG*@20. Neural embeddings are based on PV (left) and BoWE (right). Bold values show the cases where the neural embedding coherence measures lead to improve the performance of hard queries over all ranking functions.

			Statistical coherency					Statistical coherency				
	Ranking function	*modal*	*description*	*schema*	*row*	*column*	*cell*	*description*	*schema*	*row*	*column*	*cell*
Embedding coherency	BRM	*description*	0.44	1.03	**4.53**	−2.83	0.33	−0.009	−0.41	**1.93**	−1.13	1.23
	DSRMM		−5.38	−2.94	**6.71**	4.38	1.72	−6.94	−1.67	**9.52**	−3.61	−6.83
	STR		−1.89	−0.82	**8.12**	2.96	−4.57	−3.16	−4.18	**3.83**	2.001	−0.90
	LTR		−1.85	3.34	**18.12**	6.12	−2.67	4.09	4.83	**16.65**	2.17	−2.65
	BRM	*schema*	−1.31	−0.39	−1.01	−4.33	−4.11	−1.67	−1.97	−1.18	−4.30	−0.55
	DSRMM		−8.07	1.83	4.30	5.60	−5.73	−6.57	2.01	9.79	−0.91	−9.39
	STR		−8.09	−4.84	−4.75	−7.79	−14.44	−7.13	−4.06	1.59	1.09	−5.21
	LTR		0.86	8.61	16.65	5.87	3.08	−0.59	14.74	7.82	0.62	−4.39
	BRM	*row*	−0.87	**0.14**	**2.66**	−1.62	1.34	−0.09	−0.99	1.75	−3.52	0.96
	DSRMM		−10.69	**0.60**	**3.52**	3.25	−0.05	−10.76	−4.42	5.46	−1.06	−3.14
	STR		−5.51	**2.83**	**3.70**	2.96	−1.34	−3.09	3.87	−0.22	4.50	0.93
	LTR		−2.05	**2.50**	**16.13**	6.77	−0.34	−2.007	2.13	13.88	2.53	−1.83
	BRM	*column*	5.60	2.35	**12.25**	−1.62	3.82	5.10	−0.33	**11.65**	−2.99	2.39
	DSRMM		7.21	8.35	**14.64**	3.36	5.23	4.38	−2.08	**14.45**	−4.14	−5.52
	STR		1.59	−1.67	**5.36**	2.80	9.16	2.86	−3.19	**3.13**	1.51	0.06
	LTR		−3.60	0.24	**12.95**	8.06	−0.67	−4.52	1.07	**4.52**	5.13	−2.37
	BRM	*cell*	−3.57	−1.37	−1.21	−5.13	−2.16	−1.004	−1.16	2.01	−4.26	−1.78
	DSRMM		−12.22	−6.19	−2.97	−2.86	−10.56	−9.71	−5.12	4.38	−4.97	−9.52
	STR		−11.001	−3.22	1.38	−5.89	−5.19	−8.22	−7.35	−0.33	−5.49	−5.15
	LTR		−4.90	0.20	6.26	1.94	−0.09	−5.95	0.70	6.89	2.65	−0.61

Table 3
Percentage improvement of the deep contextual coherence measures over hard queries in terms of *NDCG*@20. The aggregation function is *maximum* (left) and *multiplication* (right). Bold values show the cases where the deep contextual coherence measures lead to improve the performance of hard queries over all ranking functions.

			Statistical coherency					Statistical coherency				
	Ranking function	*modal*	*description*	*schema*	*row*	*column*	*cell*	*description*	*schema*	*row*	*column*	*cell*
Deep contextual coherency	BRM	*table*	−0.22	−0.42	**2.26**	−3.83	1.36	1.56	**1.74**	**3.63**	0.15	2.68
	DSRMM		−5.80	3.48	**3.51**	−1.34	−1.98	−7.28	**1.56**	**3.27**	−2.09	−7.32
	STR		−0.33	0.79	**0.16**	1.66	1.48	4.31	**1.63**	**6.98**	2.94	3.29
	LTR		−1.56	5.48	**16.58**	5.40	3.52	1.77	**7.80**	**12.61**	6.79	0.15
	BRM	*schema*	−0.60	−0.55	**1.97**	−1.91	1.25	0.46	−0.81	**4.77**	−4.37	−2.11
	DSRMM		−6.07	3.40	**4.23**	−1.91	−2.82	−2.42	0.74	**15.18**	−0.37	−0.15
	STR		1.03	0.77	**1.57**	2.43	1.40	3.46	−0.33	**21.47**	1.30	−0.69
	LTR		−1.14	5.70	**15.56**	6.40	2.96	4.72	19.05	**21.38**	13.78	2.27
	BRM	*row*	−0.19	−0.20	**2.23**	−4.07	1.23	−2.62	−1.45	**0.39**	−5.90	−0.44
	DSRMM		−6.59	3.40	**4.19**	−1.85	−1.45	−12.79	−6.59	**1.16**	−5.95	−12.20
	STR		0.15	0.69	**0.09**	1.74	1.48	−1.96	−2.96	**1.38**	−1.78	−5.02
	LTR		−1.26	5.82	**15.50**	6.41	3.56	0.70	9.75	**30.01**	8.12	11.64
	BRM	*column*	−0.18	−0.17	2.10	−4.07	1.26	1.02	−1.45	**4.53**	−4.68	−1.70
	DSRMM		−8.01	3.10	3.29	−2.23	−5.69	−12.85	−7.65	**3.38**	−4.34	−8.13
	STR		−0.04	0.85	−0.001	2.38	−0.11	0.61	−2.79	**6.63**	−5.16	0.46
	LTR		−1.42	6.66	15.50	6.44	3.59	0.33	10.53	**23.37**	11.67	4.24
	BRM	*cell*	−0.17	−0.42	2.23	−4.07	1.26	−2.11	−1.89	**0.84**	−5.22	−1.02
	DSRMM		−7.04	3.40	4.23	−3.46	−5.61	−13.38	−8.66	**5.02**	−7.91	−12.78
	STR		−0.24	0.75	−0.11	2.33	1.47	−5.39	−4.03	**4.45**	−2.33	−4.54
	LTR		−1.56	6.01	15.46	6.49	3.04	0.24	8.47	**29.55**	14.36	13.90

ity is chosen to measure coherence, the final impact on ranking performance can vary. We find that measuring coherency based on the column modality for neural embedding coherency leads to the highest performance improvement over all four baselines. For instance, when using the PV neural representation, a column-based modality shows 12.25%, 14.64%, 5.36%, and 12.95% improvement over the BRM, DSRMM, STR, and LTR baselines, respectively. Similarly, when BoWE neural embeddings are used, an improvement of 11.65%, 14.45%, 3.13%, and 4.52% are observed on the same baselines, respectively.

Furthermore, Table 3 reports our findings based on deep contextual embeddings. We have similar observations to the neural embeddings when using deep contextual embeddings. Our results show that the column modality shows improvements over the baselines. More concretely, when considering the column modality based on the maximum aggregation

Table 4
Percentage improvement of the topical coherence measure over hard queries in terms of *NDCG@20*. Bold values show the cases where the topical coherence measure leads to improve the performance of hard queries over all ranking functions.

		Statistical coherency				
	Ranking function	*description*	*schema*	*row*	*column*	*cell*
Topical coherency	BRM	8.74	**5.33**	**9.96**	0.66	7.81
	DSRMM	−10.88	**2.51**	**7.002**	−6.44	−4.76
	STR	2.36	**1.82**	**5.68**	−3.97	−4.18
	LTR	2.96	**5.60**	**14.64**	5.23	−0.92

function, we observe 2.10%, 3.29%, and 15.5% improvements over BRM, DSRMM, and LTR, respectively, but STR is not improved. Moreover, when employing the multiplication aggregation function, we observe 4.53%, 3.38%, 6.63%, and 23.37% improvements over BRM, DSRMM, STR, and LTR, respectively. This shows that the aggregation function has a notable impact on the performance of the deep contextualized coherence measures where the multiplication function is more effective when compared with the maximum function.

Finally, in Table 4, we report the performance of the topical coherency measure. Given topic coherency compares the topic distributions of the table content with the table description, different table modalities are not considered for this measure. Overall, we find that topical coherency also shows improvements of 9.96%, 7.002%, 5.68%, and 14.64% over all four baselines, namely BRM, DSRMM, STR, and LTR, respectively. In terms of the statistical coherency measure, we find that row modality has shown the best performance consistently across all baselines. We also find that when coherency is measured based on semantic representations or topic information, the column modality shows the best performance.

To determine which measure of table coherence shows the best, most consistent performance, we consider the *row modality* for statistical coherence and *column modality*for the other coherency metrics as identified above. In such a context, when considering the BRM baseline, the best consistent improvements are 12.25%, 4.53%, and 9.96% based on the neural embedding, deep contextual embedding, and topical coherence in conjunction with statistical coherence, respectively. For the DSRMM baseline, the best consistent performance improvements are 14.64%, 3.38%, and 7.002%, respectively. Given STR as the baseline ranking function, the improvements are 5.36%, 6.63%, and 5.68%, respectively. Finally, for LTR baseline, there are 12.95%, 23.37%, and 14.64% improvements over hard queries. We observe that the best coherence measure for BRM and DSRMM methods are statistical and neural embedding coherency metrics while STR and LTR are best improved through statistical and deep contextual embedding coherency metrics while neural embedding and topical coherence measures are also highly competitive.

Finding 1: We summarize our findings on the coherence measures as follows:

- The application of different coherence measures on different modalities does not always lead to improvement on the retrieval of hard queries in ad hoc table retrieval.
- The statistical coherence measure based on the row modality and semantic and topical coherence measures on the column modality show consistent improved performance over all four state-of-the-art methods.

Now, we further consider the impact of table interpretability measures on the performance of hard queries in table retrieval. We view interpretability as the extent of pertinent information available to extend table content to make it more understandable in the context of a user query. We have suggested it is possible to extend a table either through the text that surrounds the table or by the content accessible through the links made available in the table. As outlined in Table 1, we introduce statistical and semantic table interpretability measures. Table 5 provides an overview of the impact of the interpolation of statistical and semantic interpretability metrics on four baseline retrieval methods. In this table, the first column shows the interpretability measures where the surrounding content of the table is used to extend the table while the second shows the case when content from links provided in the table are used to extend the table. When comparing the two columns, we find that a link-based table extension provides substantially better performance when compared with the surrounding text-based extension method. In other words, content linked directly from within the table is more helpful in interpreting the table content when compared with the content surrounding the table. Furthermore, within a link-based extension of the table, the variation of the interpretability metric that computes the likelihood of the extension of the table being generated from the query shows the best performance.

We observe that while BRM and STR improved by 2.34% and 6.96%, respectively, the DSRMM and LTR saw an overwhelming higher performance boost as a result of the interpretability measures. We believe this could be because DSRMM limits itself to the first fifty tokens from the table description, the first thirty tokens from the table schema, and twenty tokens from table rows and columns [12]. Therefore, DSRMM loses a lot of information about the table, while the table extension approach in the table interpretability measures compensates for this lack of information.

Table 5
Percentage improvement of the proposed table interpretability measures over hard queries in terms of *NDCG*@20. The cases where the interpretability measures provide the improvement over all ranking functions are bold.

		Statistical interpretability	
	Ranking function	Surrounding text-based	Link-based
Embedding interpretability (PV)	BRM	3.05	**1.66**
	DSRMM	25.46	**13.63**
	STR	-6.39	**6.96**
	LTR	22.33	**28.99**
Embedding interpretability (BoWE)	BRM	2.46	**2.07**
	DSRMM	26.28	**17.93**
	STR	-4.85	**5.35**
	LTR	19.95	**25.68**
Deep contextual interpretability	BRM	2.75	**2.34**
	DSRMM	26.38	**17.20**
	STR	-4.37	**5.57**
	LTR	20.64	**24.20**

Finding 2: We summarize our findings on the interpretability measures as follows:

- The information accessible through links provided within tables provide better context for extending tables, which can lead to better understandability of the table, and hence more effective retrieval, when compared with when the surrounding content of the table are used.
- The interpolation of statistical and semantic interpretability measures lead to consistent positive improvement of all baseline methods, but the improvements are more notable on the DSRMM and LTR baselines.

The third proposed class of measures relates to table exactness, which measures how much the content in table modalities maximally align with the information need of the user. Tables 6–8 report the results of interpolating statistical exactness measures with semantic exactness measures into the baseline retrievers. In Table 6, a BoWE approach is used to operationalize the semantic exactness measure. We find that, regardless of the adopted aggregation function, the row and cell modalities are quite effective representations for the statistical exactness measure. We further observe that the consid-

Table 6
Percentage improvement of the proposed BoWE exactness measures over hard queries in terms of *NDCG*@20 based on *summation* (left) and *maximum* (right) aggregation functions. Bold values indicate the cases where the BoWE exactness measures lead to improve the performance of hard queries over all ranking functions.

			Statistical exactness					Statistical exactness				
		Ranking function	*description*	*schema*	*row*	*column*	*cell*	*description*	*schema*	*row*	*column*	*cell*
Embedding exactness	*description*	BRM	**2.99**	1.51	**12.75**	**2.05**	**9.45**	−0.24	−0.73	**4.84**	−0.13	**6.44**
		DSRMM	**2.12**	−2.04	**25.62**	**11.35**	**23.72**	−1.21	−3.53	**28.19**	9.51	**28.20**
		STR	**6.56**	10.36	**34.98**	**12.58**	**32.26**	−2.47	10.35	**18.42**	8.96	**24.96**
		LTR	**7.23**	10.69	**25.62**	**23.57**	**26.69**	9.86	21.88	**25.08**	21.14	**27.02**
	schema	BRM	−0.29	−0.98	**2.22**	−1.42	**2.26**	0.69	−0.96	**3.66**	−1.21	**3.54**
		DSRMM	2.44	4.54	**35.66**	9.77	**34.68**	−1.79	11.39	**29.16**	13.37	**40.08**
		STR	−0.10	4.97	**23.27**	10.77	**29.52**	17.09	13.63	**26.91**	19.52	**36.34**
		LTR	6.85	10.09	**21.54**	24.75	**23.82**	25.32	18.89	**22.38**	21.01	**22.28**
	row	BRM	−0.05	−1.05	**3.76**	−1.34	**4.69**	**2.50**	**0.50**	**11.29**	−0.33	**13.38**
		DSRMM	8.20	0.04	**23.66**	11.07	**25.41**	**8.04**	**1.15**	**20.43**	3.12	**23.63**
		STR	−0.26	11.34	**20.63**	15.88	**25.16**	**0.04**	**11.29**	**17.73**	9.50	**26.38**
		STR	3.33	2.96	**17.34**	17.73	**17.70**	**18.06**	**28.05**	**24.78**	21.003	**25.62**
	column	BRM	1.14	**0.10**	**3.09**	**0.49**	**3.82**	2.57	**0.59**	**11.81**	−0.26	**13.36**
		DSRMM	4.36	**6.66**	**12.27**	**7.66**	**12.63**	11.71	**1.56**	**20.60**	1.35	**24.26**
		STR	−8.25	**1.48**	**14.24**	**4.65**	**13.13**	−0.10	**12.29**	**17.98**	8.94	**25.98**
		LTR	3.51	**12.97**	**15.54**	**14.63**	**15.01**	17.96	**27.74**	**24.85**	20.48	**25.55**
	cell	BRM	−2.23	−1.92	**3.23**	−1.54	**3.25**	**2.67**	**0.36**	**11.12**	−0.27	**13.36**
		DSRMM	7.21	6.80	**31.60**	19.87	**33.50**	**12.03**	**1.56**	**20.43**	1.59	**24.26**
		STR	−8.54	7.31	**12.76**	5.27	**13.34**	**0.09**	**12.26**	**20.11**	8.97	**26.02**
		LTR	8.53	7.84	**19.78**	11.53	**18.95**	**17.96**	**27.86**	**24.37**	20.45	**25.37**

Table 7
Percentage improvement of the proposed PV exactness measure over hard queries in terms of *NDCG*@20. Bold values indicate the cases where the PV exactness measure leads to performance improvement of hard queries over all ranking functions.

			Statistical exactness				
		Ranking function	*description*	*schema*	*row*	*column*	*cell*
Embedding exactness (PV)	*description*	BRM	0.99	−0.35	**4.77**	**1.10**	**6.39**
		DSRMM	−3.03	−3.29	**29.19**	**9.74**	**34.39**
		STR	−3.08	8.38	**18.02**	**8.13**	**23.88**
		LTR	9.50	14.28	**23.82**	**23.47**	**29.57**
	schema	BRM	**1.19**	−0.23	**2.17**	−1.04	**3.56**
		DSRMM	**1.73**	0.98	**31.26**	11.07	**29.53**
		STR	**7.85**	13.77	**30.51**	19.71	**33.61**
		LTR	**15.46**	18.61	**23.01**	24.14	**24.97**
	row	BRM	**2.37**	**0.88**	**11.13**	**0.11**	**13.32**
		DSRMM	**0.87**	**2.33**	**23.85**	**5.82**	**32.48**
		STR	**3.73**	**14.83**	**16.31**	**11.15**	**25.72**
		LTR	**23.67**	**34.11**	**23.35**	**20.76**	**25.38**
	column	BRM	**1.08**	**0.30**	**11.81**	−0.27	**7.10**
		DSRMM	**4.64**	**0.15**	**23.49**	2.34	**26.32**
		STR	**0.001**	**10.43**	**18.86**	8.60	**25.90**
		LTR	**26.62**	**32.63**	**24.22**	20.15	**24.82**
	cell	BRM	**2.19**	**0.44**	**11.53**	**0.53**	**13.43**
		DSRMM	**2.67**	**0.18**	**22.16**	**2.79**	**27.19**
		STR	**1.91**	**14.69**	**15.78**	**8.70**	**24.71**
		LTR	**22.50**	**34.67**	**22.64**	**19.90**	**24.46**

Table 8
Percentage improvement of the proposed deep contextualized exactness over hard queries in terms of *NDCG*@20. Bold values indicate the cases where the deep contextualized exactness measure leads to improve the performance of hard queries over all ranking functions.

			Statistical exactness				
		Ranking function	*description*	*schema*	*row*	*column*	*cell*
Deep contextual exactness	*description*	BRM	1.17	−0.89	**6.12**	**0.15**	**5.69**
		DSRMM	−2.60	−2.75	**24.52**	**7.09**	**25.20**
		STR	−3.39	11.21	**18.30**	**9.06**	**26.08**
		LTR	8.77	18.77	**26.07**	**21.48**	**29.23**
	schema	BRM	1.03	−0.68	**6.14**	−0.14	**7.48**
		DSRMM	−1.90	−0.54	**23.67**	4.06	**27.86**
		STR	−3.61	11.02	**16.25**	8.97	**26.70**
		LTR	8.48	17.95	**24.26**	23.27	**26.75**
	row	BRM	2.45	0.21	**5.62**	**0.66**	**7.78**
		DSRMM	2.11	−2.61	**24.18**	**3.91**	**26.22**
		STR	−0.20	11.85	**20.40**	**9.23**	**25.99**
		LTR	12.50	19.24	**25.39**	**21.61**	**27.96**
	column	BRM	1.93	−0.61	**5.43**	**0.47**	**7.75**
		DSRMM	−7.23	−2.92	**23.73**	**3.41**	**26.21**
		STR	0.22	11.88	**20.41**	**9.27**	**26.53**
		LTR	8.46	19.92	**26.09**	**22.45**	**27.11**
	cell	BRM	0.93	−0.91	**5.12**	**0.07**	**7.48**
		DSRMM	−2.18	−3.50	**23.30**	**3.67**	**23.82**
		STR	−0.77	12.03	**21.30**	**9.21**	**25.90**
		LTR	8.36	17.94	**24.57**	**20.74**	**26.40**

eration of the information in the schema modality is the most impactful for capturing BoWE semantic exactness. It is also possible to see that while both aggregation methods are quite effective at improving the performance of the baseline retrievers, the maximum aggregation function shows better overall performance improvements over all four baselines.

We also measure semantic exactness through PV as reported in Table 7. Like the BoWE, both cell and row modalities are effective when measuring statistical exactness in conjunction with PV-based semantic exactness. In contrast, however, the best modality for the PV-based semantic exactness measure depends on the baseline retriever. For both DSRMM and STR, the schema modality is the most efficient representation, which is consistent with the BoWE approach. For BRM and LTR, the cell and description modalities respectively show the best performance. However, the modality that shows significant and consistent improvement over all baselines in the PV-based exactness measure is the cell modality, which reports 13.43%, 27.19%, 24.71%, and 24.46% improvements over BRM, DSRMM, STR and LTR, respectively.

Finally, Table 8 reports the performance of the deep contextualized embedding-based exactness measures. Again, like the previous exactness measures, cell and row modalities are the best for the statistical exactness measure. Also like the BoWE

approach, the schema modality shows the best performance improvement based on the deep contextualized representation. This is equivalent to 7.48%, 27.86%,26.70%, and 26.75% over BRM, DSRMM, STR, and LTR baselines, respectively.

We further explore the most notable impact by the proposed exactness measures on the performance of the baseline retrieval methods. We already discussed that the cell modality is the strongest representation for the statistical exactness measure while row and schema modalities are the stronger representations for the semantic exactness measures. Based on these modalities, we find that BRM can be improved by 13.38%, 13.32%, and 7.78% based on BoWE, PV, and deep contextualized embeddings, respectively. For BRM, both BoWE and PV formalizations show a similar performance. For the DSRMM method, we observe 40.08%, 29.53%, and 27.86%, respectively. Also, STR observes 36.34%, 33.61%, and 26.70% improvements based on BoWE and PV and deep contextualized embeddings. Finally, for the LTR baseline, they are 22.28%, 24.97%, and 26.75% improvements over hard queries. Based on these results, we believe that the BoWE representation shows a stable and consistent representation for the exactness measure that shows notable improvements on all baselines.

Finding 3: We summarize our findings on the exactness measures as follows:

- In terms of the suitability of table modalities, we find the cell modality shows the best performance on the statistical exactness measure while the row and schema modalities show the highest impact for the semantic exactness measures.
- From among the three types of exactness measures, the BoWE semantic representation shows the most consistent, stable, and notable improvement over the baseline retrievers and for all the exactness measures.

Now, we recognize that the performance of each of the four baseline retrieval methods is not the same on all queries and therefore the percentage of improvements reported by each of the proposed qualitative measures are not directly comparable between the baselines. In other words, the percentage of improvements are only meaningful within the context of each retrieval method. As such, and to see how the proposed qualitative measures can impact the performance of the baseline retrieval method, we report the performance of the qualitative methods separately for each baseline retrieval method in Tables 9–12.

For the BRM retrieval method, we find that the row and cell modalities exhibit the highest performance on the statistical exactness measure while the row modality shows the best performance on the proposed statistical and topical coherence measures. In the context of the interpretability measure, both surrounding text-based and link-based methods show improvement over the baseline retriever. Overall, for the BRM retriever, we find that the best performance improvements happen by the proposed qualitative measures when the exactness measures are applied through row and cell modalities. Table 10.

In the context of the DSRMM method, we observe that, like the BRM method, both row and cell modalities perform well for the exactness measure and the row modality is similarly quite appropriate on the coherence measure. Likewise, both types of the interpretability measure show improvements over the performance of the DSRMM method. Overall, we observe that the highest and most consistent performance improvements over the DSRMM method are reported when the cell modality is adopted in the context of statistical exactness. Table 11.

With regards to the STR retrieval method, similar to the previous two retrieval methods, the row and cell modalities perform quite well on the exactness measure and the row modality for the coherence measure. However, unlike the other two retrieval methods, the interpretability measure only shows performance improvement when applied through the link-based variation. Overall, we find that while the coherence and interpretability measures show performance improvements, the highest and the most consistent improvements over the STR method are observed through the exactness measure and especially using the row and cell modalities.

Finally, when applying the proposed qualitative measures on the LTR retrieval method, we find that the base retriever improves when interpolated with most of the proposed qualitative methods (except for a few cases such as the cell and description modalities on coherence and cell modality on topical coherence). However, we observe that, like the other four baseline retrieval methods, the highest and most consistent improvements are observed through the proposed exactness measure.

Finding 4. The proposed qualitative exactness measure, especially when applied through the cell and row modalities, shows the highest and most consistent performance improvements across all four baseline retrieval methods. This means that the exactness measure is the most effective across all four baseline retrieval methods when interpolated with each of them separately

Table 9
Percentage improvement of the proposed quality measures over hard queries for BRM ranking function in terms of NDCG@20.

		Embedding exactness (*summation*)					Embedding exactness (*maximum*)				
Statistical exactness		*description/ table*	*schema*	*row*	*column*	*cell*	*description/ table*	*schema*	*row*	*column*	*cell*
	description	2.99	-0.29	-0.05	1.14	-2.23	-0.24	0.69	2.50	2.57	2.67
	schema	1.51	-0.98	-1.05	0.10	-1.92	-0.73	-0.96	0.50	0.59	0.36
	row	12.75	2.22	3.76	3.09	3.23	4.84	3.66	11.29	11.81	11.12
	column	2.05	-1.42	-1.34	0.49	-1.54	-0.13	-1.21	-0.33	-0.26	-0.27
	cell	9.45	2.26	4.69	3.82	3.25	6.44	3.54	13.38	13.36	13.36
		Embedding exactness (PV)					Deep contextual exactness				
	description	0.99	1.19	2.37	1.08	2.19	1.17	1.03	2.45	1.93	0.93
	schema	-0.35	-0.23	0.88	0.30	0.44	-0.89	-0.68	0.21	-0.61	-0.91
	row	4.77	2.17	11.13	11.81	11.53	6.12	6.14	5.62	5.43	5.12
	column	1.10	-1.04	0.11	-0.27	0.53	0.15	-0.14	0.66	0.47	0.07
	cell	6.39	3.56	13.32	7.10	13.43	5.69	7.48	7.78	7.75	7.48
Statistical coherency		Embedding coherency (PV)					Embedding coherency (BoWE)				
	description	0.44	-1.31	-0.87	5.60	-3.57	-0.009	-1.67	-0.09	5.10	-1.004
	schema	1.03	-0.39	0.14	2.35	-1.37	-0.41	-1.97	-0.99	-0.33	-1.16
	row	4.53	-1.01	2.66	12.25	-1.21	1.93	-1.18	1.75	11.65	2.01
	column	-2.83	-4.33	-1.62	-1.62	-5.13	-1.13	-4.30	-3.52	-2.99	-4.26
	cell	0.33	-4.11	1.34	3.82	-2.16	1.23	-0.55	0.96	2.39	-1.78
		Deep contextual coherency (*maximum*)					Deep contextual coherency (*multiplication*)				
	description	-0.22	-0.60	-0.19	-0.18	-0.17	1.56	0.46	-2.62	1.02	-2.11
	schema	-0.42	-0.55	-0.20	-0.17	-0.42	1.74	-0.81	-1.45	-1.45	-1.89
	row	2.26	1.97	2.23	2.10	2.23	3.63	4.77	0.39	4.53	0.84
	column	-3.83	-1.91	-4.07	-4.07	-4.07	0.15	-4.37	-5.90	-4.68	-5.22
	cell	1.36	1.25	1.23	1.26	1.26	2.68	-2.11	-0.44	-1.70	-1.02
		Topical coherency									
	description	8.74									
	schema	5.33									
	row	9.96									
	column	0.66									
	cell	7.81									
Interpretability		Embedding interpretability (PV)				Embedding interpretability (BoWE)			Deep contextual interpretability		
	Surrounding text-based	3.05				2.46			2.75		
	Link-based	1.66				2.07			2.34		

Let us explore the reason why the exactness measure has a higher likelihood of improving hard queries through a sample hard query 'eu countries year joined'. For this query, the relevant table is placed at rank ten by the best performing baseline, namely STR. Fig. 6 shows a snapshot of the relevant table. The rank of the relevant table improves to rank eight only after the application of the exactness measure but does not change when the other measures are applied. We believe this is a good query to illustrate the impact of the exactness measure because (1) the query is difficult to address by all baselines and the best baseline retrieves the relevant document at rank 10, and (2) it is only the exactness measure that can improve this query and the other measures have no impact.

When reviewing this relevant table, we find that the table lists European Union countries and provides various types of information in different columns. Several of the columns in the table are not related to our query. These columns include 'Compulsory/optional', 'Cost', 'Validity', 'Issuing authority', and 'Latest version'. For this reason, any retrieval method or, for that matter, any of our qualitative measures that consider all the information in such a table are prone to misjudging the relevance of the table to the query. In this case, when considering the whole table, this relevant table would need to be ranked quite low. However, the exactness measure effectively addresses this problem by focusing on the most relevant modality of the table to the query (in this case, the first column of the table) and considers it to measure the association between the query and the table. The exactness measure approach allows the retrieval method to focus only on the appropriate parts of the table when measuring relevance, hence leading to reasonable performance.

In contrast, while the exactness measure focuses on pinpointing relevant information within a table, the other measures could, in some cases, be prone to the introduction of additional noise to the relevance measurement process. For instance, in this specific case, there are many links in the table such as footnotes referring to websites such as Belgian government's Identity and Civic Affairs website that are not related to the query topic. These links not only do not add any relevant information but could also introduce additional irrelevant information that would mislead the process of computing table-query

Table 10
Percentage improvement of the proposed quality measures over hard queries for DSRMM ranking function in terms of NDCG@20.

		Embedding exactness (*summation*)					Embedding exactness (*maximum*)				
		description/ table	*schema*	*row*	*column*	*cell*	*description/ table*	*schema*	*row*	*column*	*cell*
Statistical exactness	*description*	2.12	2.44	8.20	4.36	7.21	-1.21	-1.79	8.04	11.71	12.03
	schema	-2.04	4.54	0.04	6.66	6.80	-3.53	11.39	1.15	1.56	1.56
	row	25.62	35.66	23.66	12.27	31.60	28.19	29.16	20.43	20.60	20.43
	column	11.35	9.77	11.07	7.66	19.87	9.51	13.37	3.12	1.35	1.59
	cell	23.72	34.68	25.41	12.63	33.50	28.20	40.08	23.63	24.26	24.26
		Embedding exactness (PV)					Deep contextual exactness				
	description	-3.03	1.73	0.87	4.64	2.67	-2.60	-1.90	2.11	-7.23	-2.18
	schema	-3.29	0.98	2.33	0.15	0.18	-2.75	-0.54	-2.61	-2.92	-3.50
	row	29.19	31.26	23.85	23.49	22.16	24.52	23.67	24.18	23.73	23.30
	column	9.74	11.07	5.82	2.34	2.79	7.09	4.06	3.91	3.41	3.67
	cell	34.39	29.53	32.48	26.32	27.19	25.20	27.86	26.22	26.21	23.82
Statistical coherency		Embedding coherency (PV)					Embedding coherency (BoWE)				
	description	-5.38	-8.07	-10.69	7.21	-12.22	-6.94	-6.57	-10.76	4.38	-9.71
	schema	-2.94	1.83	0.60	8.35	-6.19	-1.67	2.01	-4.42	-2.08	-5.12
	row	6.71	4.30	3.52	14.64	-2.97	9.52	9.79	5.46	14.45	4.38
	column	4.38	5.60	3.25	3.36	-2.86	-3.61	-0.91	-1.06	-4.14	-4.97
	cell	1.72	-5.73	-0.05	5.23	-10.56	-6.83	-9.39	-3.14	-5.52	-9.52
		Deep contextual coherency (*maximum*)					Deep contextual coherency (*multiplication*)				
	description	-5.80	-6.07	-6.59	-8.01	-7.04	-7.28	-2.42	-12.79	-12.85	-13.38
	schema	3.48	3.40	3.40	3.10	3.40	1.56	0.74	-6.59	-7.65	-8.66
	row	3.51	4.23	4.19	3.29	4.23	3.27	15.18	1.16	3.38	5.02
	column	-1.34	-1.91	-1.85	-2.23	-3.46	-2.09	-0.37	-5.95	-4.34	-7.91
	cell	-1.98	-2.82	-1.45	-5.69	-5.61	-7.32	-0.15	-12.20	-8.13	-12.78
		Topical coherency									
	description	-10.88									
	schema	2.51									
	row	7.002									
	column	-6.44									
	cell	-4.76									
Interpretability		Embedding interpretability (PV)			Embedding interpretability (BoWE)			Deep contextual interpretability			
	Surrounding text-based	25.46			26.28			26.38			
	Link-based	13.63			17.93			17.20			

relevance. As such, the interpretability measure, which adds contextual information to each table through links, can lose its impact when there are too many unrelated links to the query in a table.

Furthermore, the coherency measure advocates for the desirability of tables where the content in each of the table's modalities are consistent with each other. We believe that coherent tables are more desirable for the users since, if relevant to the query, all the information in the table will be focused on a similar topic and hence would be easier for the user to understand and interpret. However, in practice, there are tables containing the relevant information that can address the information needs of the user. However, these also include other pieces of information that may not be directly related to the query. The expectation imposed by the coherency measure will negatively impact such tables, including the table shown in Fig. 6. One could argue that, ideally, a relevant and coherent table would be the most desirable table for a user query. However, in practice, one might not be able to find a relevant and fully coherent table. As such, a measure such as exactness that considers the specific relevant modalities of a table to the query instead of requiring the whole table to be consistent shows better overall performance.

5.5.2. RQ2. The impact of the qualitative measures on table retrieval

In the second research question, we aim to explore the impact of the proposed qualitative measures on the table retrieval task on the hard queries of the benchmark dataset. Based on the observations summarized in Findings 1–4, we chose the best variation of each qualitative metric as discussed in RQ1. Table 13 summarizes these consistent variations of each qualitative measure and reports their performance.

As observed in the table, the proposed qualitative measures have effectively improved the performance of all baseline methods over the hard queries. We make several observations based on the results in Table 13. In terms of performance,

Table 11
Percentage improvement of the proposed quality measures over hard queries for STR ranking function in terms of NDCG@20.

		Embedding exactness (*summation*)					Embedding exactness (*maximum*)				
		description/ table	*schema*	*row*	*column*	*cell*	*description/ table*	*schema*	*row*	*column*	*cell*
Statistical exactness	*description*	6.56	-0.10	-0.26	-8.25	-8.54	-2.47	17.09	0.04	-0.10	0.09
	schema	10.36	4.97	11.34	1.48	7.31	10.35	13.63	11.29	12.29	12.26
	row	34.98	23.27	20.63	14.24	12.76	18.42	26.91	17.73	17.98	20.11
	column	12.58	10.77	15.88	4.65	5.27	8.96	19.52	9.50	8.94	8.97
	cell	32.26	29.52	25.16	13.13	13.34	24.96	36.34	26.38	25.98	26.02
		Embedding exactness (PV)					Deep contextual exactness				
	description	-3.08	7.85	3.73	0.001	1.91	-3.39	-3.61	-0.20	0.22	-0.77
	schema	8.38	13.77	14.83	10.43	14.69	11.21	11.02	11.85	11.88	12.03
	row	18.02	30.51	16.31	18.86	15.78	18.30	16.25	20.40	20.41	21.30
	column	8.13	19.71	11.15	8.60	8.70	9.06	8.97	9.23	9.27	9.21
	cell	23.88	33.61	25.72	25.90	24.71	26.08	26.70	25.99	26.53	25.90
Statistical coherency		Embedding coherency (PV)					Embedding coherency (BoWE)				
	description	-1.89	-8.09	-5.51	1.59	-11.001	-3.16	-7.13	-3.09	2.86	-8.22
	schema	-0.82	-4.84	2.83	-1.67	-3.22	-4.18	-4.06	3.87	-3.19	-7.35
	row	8.12	-4.75	3.70	5.36	1.38	3.83	1.59	-0.22	3.13	-0.33
	column	2.96	-7.79	2.96	2.80	-5.89	2.001	1.09	4.50	1.51	-5.49
	cell	-4.57	-14.44	-1.34	9.16	-5.19	-0.90	-5.21	0.93	0.06	-5.15
		Deep contextual coherency (*maximum*)					Deep contextual coherency (*multiplication*)				
	description	-0.33	1.03	0.15	-0.04	-0.24	4.31	3.46	-1.96	0.61	-5.39
	schema	0.79	0.77	0.69	0.85	0.75	1.63	-0.33	-2.96	-2.79	-4.03
	row	0.16	1.57	0.09	-0.001	-0.11	6.98	21.47	1.38	6.63	4.45
	column	1.66	2.43	1.74	2.38	2.33	2.94	1.30	-1.78	-5.16	-2.33
	cell	1.48	1.40	1.48	-0.11	1.47	3.29	-0.69	-5.02	0.46	-4.54
		Topical coherency									
	description	2.36									
	schema	1.82									
	row	5.68									
	column	-3.97									
	cell	-4.18									
Interpretability		Embedding interpretability (PV)			Embedding interpretability (BoWE)				Deep contextual interpretability		
	Surrounding text-based	-6.39			-4.85				-4.37		
	Link-based	6.96			5.35				5.57		

the BRM method has the least impact through the interpretability measures, while both coherency and exactness measures exhibit improvements of 12–13%. The DSRMM method see the greatest improvement based on the exactness measures while the coherency and interpretability measures show similar degrees of impact of 15–17%. Furthermore, the STR and LTR methods see the most significant degree of improvement because of being interpolated with coherency and exactness measures. STR method, unlike LTR, does not see as much improvement based on the interpretability measures.

Finding 5: We summarize our findings on the overall performance of the proposed measures as follows:

- The interpretability measures show the least degree of impact on the performance of the baselines, especially the BRM and STR baselines. In contrast, the exactness measures show the strongest performance improvement on all baselines.
- From the perspective of the baselines and comparatively speaking, the BRM method sees the least degree of improvement because of the proposed qualitative measures where its best improvement is obtained by the exactness method. In contrast, the DSRMM, STR, and LTR methods improve much more significantly by applying the coherency and exactness measures.

Given our proposed qualitative measures are used within the context of a re-ranking framework, as defined in Eq. 1, we are interested in comparing the performance of our proposed re-ranking framework with the state-of-the-art methods that perform re-ranking. We compare our work with the Projection-based Table Relevance Model (PTRM) [17] that builds on top

Table 12
Percentage improvement of the proposed quality measures over hard queries for LTR ranking function in terms of NDCG@20.

		Embedding exactness (*summation*)					Embedding exactness (*maximum*)				
		description/ table	*schema*	*row*	*column*	*cell*	*description/ table*	*schema*	*row*	*column*	*cell*
Statistical exactness	*description*	7.23	6.85	3.33	3.51	8.53	9.86	25.32	18.06	17.96	17.96
	schema	10.69	10.09	2.96	12.97	7.84	21.88	18.89	28.05	27.74	27.86
	row	25.62	21.54	17.34	15.54	19.78	25.08	22.38	24.78	24.85	24.37
	column	23.57	24.75	17.73	14.63	11.53	21.14	21.01	21.003	20.48	20.45
	cell	26.69	23.82	17.70	15.01	18.95	27.02	22.28	25.62	25.55	25.37
		Embedding exactness (PV)					Deep contextual exactness				
	description	9.50	15.46	23.67	26.62	22.50	8.77	8.48	12.50	8.46	8.36
	schema	14.28	18.61	34.11	32.63	34.67	18.77	17.95	19.24	19.92	17.94
	row	23.82	23.01	23.35	24.22	22.64	26.07	24.26	25.39	26.09	24.57
	column	23.47	24.14	20.76	20.15	19.90	21.48	23.27	21.61	22.45	20.74
	cell	29.57	24.97	25.38	24.82	24.46	29.23	26.75	27.96	27.11	26.40
Statistical coherency		Embedding coherency (PV)					Embedding coherency (BoWE)				
	description	-1.85	0.86	-2.05	-3.60	-4.90	4.09	-0.59	-2.007	-4.52	-5.95
	schema	3.34	8.61	2.50	0.24	0.20	4.83	14.74	2.13	1.07	0.70
	row	18.12	16.65	16.13	12.95	6.26	16.65	7.82	13.88	4.52	6.89
	column	6.12	5.87	6.77	8.06	1.94	2.17	0.62	2.53	5.13	2.65
	cell	-2.67	3.08	-0.34	-0.67	-0.09	-2.65	-4.39	-1.83	-2.37	-0.61
		Deep contextual coherency (*maximum*)					Deep contextual coherency (*multiplication*)				
	description	-1.56	-1.14	-1.26	-1.42	-1.56	1.77	4.72	0.70	0.33	0.24
	schema	5.48	5.70	5.82	6.66	6.01	7.80	19.05	9.75	10.53	8.47
	row	16.58	15.56	15.50	15.50	15.46	12.61	21.38	30.01	23.37	29.55
	column	5.40	6.40	6.41	6.44	6.49	6.79	13.78	8.12	11.67	14.36
	cell	3.52	2.96	3.56	3.59	3.04	0.15	2.27	11.64	4.24	13.90
		Topical coherency									
	description	2.96									
	schema	5.60									
	row	14.64									
	column	5.23									
	cell	-0.92									
Interpretability		Embedding interpretability (PV)					Embedding interpretability (BoWE)				
	Surrounding text-based	22.33					19.95				
	Link-based	28.99					25.68				

of the relevance model. This work has reported its performance on the WikiTables corpus and re-ranked the results of STR, but has not released its code or runs. Therefore, for the sake of consistency with its results, we must compare the performance of our work with this baseline based on all queries (and not hard queries as we do not have access to the performance of these methods on a per query basis). For the sake of comparison and based on Finding 5, we re-rank the results of STR, as done in [17], and report the performance of the different variations of our proposed work when interpolated with STR. Table 14 reports the results of re-ranking STR using the two different baseline methods as well as our proposed measures along with the performance of the baseline retrieval methods. We make several observations based on Table 14. First, while the re-ranking methods including the work in [17] can improve the performance of the baseline method, i.e., STR, they are can not show better performance compared with the performance of the best retrieval method, namely DSRMM. In comparison STR + PTRM [17], when compared with DSRMM on NDCG at ranks 5, 10, and 20 are 61.76 vs 64, 64.32 vs 65.7, and 69.05 vs 70.3, respectively. Second, we observe that the interpolation of our proposed qualitative methods with STR as suggested in [17] not only improves the performance of STR but also consistently shows better performance when compared with the best retrieval baselines including DSRMM (except the STR + Coherency + Exactness variation on NDCG@5). This means that each of the proposed measures individually and also in tandem, show superior performance when compared with the best retrieval baselines. Finally, given the performances reported in Table 14 are based on all queries, not just hard queries, we conclude that our proposed metrics are not only effective on hard queries but are also effective across the range of all queries in the query set.

5.5.3. RQ3. Synergistic impact of the qualitative measures

In the third research question, we would like to identify whether the proposed qualitative measures in this paper have complementary and synergistic behavior on each other. To do so, we select the best performing variation of each class of metric based on the results reported in Research Questions 1 and 2, and linearly interpolate these selected qualitative measures. We first perform the linear interpolation of each pair of metrics where the best variation of one class of qualitative measures is interpolated with the best variation of the other. This will produce three interpolated models. We then further

Fig. 6. A sample relevant table to the hard query 'eu countries year joined'.

Table 13
The impact of each of the qualitative measures on the performance of the baselines over the hard queries based on *NDCG*@20.

	Ranking function	*NDCG*@20	*NDCG*@20Δ%
Baseline performance	BRM	0.31	-
	DSRMM	0.29	-
	STR	0.41	-
	LTR	0.24	-
Coherency	BRM	0.35	12.25
	DSRMM	0.33	15.18
	STR	0.49	21.47
	LTR	0.31	27.43
Interpretability	BRM	0.32	2.34
	DSRMM	0.34	17.20
	STR	0.43	5.57
	LTR	0.29	19.01
Exactness	BRM	0.35	13.43
	DSRMM	0.41	40.08
	STR	0.55	36.38
	LTR	0.32	31.48

Table 14
Comparative analysis of our proposed re-ranking framework with state-of-the-art re-ranking methods on all queries in the WikiTables corpus.

Re-ranking Method	*NDCG*@5	*NDCG*@10	*NDCG*@20
LTR	55.27	54.56	60.31
STR	59.51	62.93	68.25
BRM	62.74	64.65	65.32
DSRMM	64.00	65.7	70.3
STR + Coherency	65.5	66.4	71.2
STR + Interpretability	67.63	68.1	73.3
STR + Exactness	64.8	65.9	70.6
STR + Coherency + Interpretability	66.5	67.3	72.1
STR + Coherency + Exactness	63.9	65.9	70.9
STR + Interpretability + Exactness	64.4	66.1	71
STR + Coherency + Interpretability + Exactness	65.9	66.7	71.5
STR + RM3	61.64	64.16	69.85
STR + PTRM	61.76	64.32	69.05

interpolate one variation of each class of qualitative measures with the others, creating one interpolation which consists of three qualitative measures. The results are reported in Table 15.

From the perspective of the BRM method, we observe this model consistently improves based on the four interpolated models. The least amount of improvement, equivalent to 6.56% over the performance of BRM, is seen when the best variations of coherence and interpretability are interpolated. The most improvement is observed when all measures are interpolated. Like BRM, DSRMM consistently improves by the various interpolated methods, however, unlike BRM, DSRMM does not experience the most improvement based on the interpolation of the three qualitative measures, but rather it observes the most improvement as a result of the interpolation of coherence and exactness. From the point of view of STR, when employing the interpolated variation including coherence and interpretability, the performance drops by 1.25%. Other than that, the other variations show increased performance. Like DSRMM, STR and LTR also experience the highest improvement on the variation that interpolates coherence and exactness. Table 16.

Table 15
Interpolation of the best variation of the qualitative measures on hard queries in the WikiTables corpus.

	Ranking function	*NDCG*@20	*NDCG*@20Δ%
Coherence + Interpretability	BRM	0.33	6.56
	DSRMM	0.31	7.34
	STR	0.40	−1.25
	LTR	0.28	17.79
Coherence + Exactness	BRM	0.33	7.37
	DSRMM	0.34	18.19
	STR	0.50	22.94
	LTR	0.31	26.48
Interpretability + Exactness	BRM	0.33	7.30
	DSRMM	0.34	15.62
	STR	0.44	8.85
	LTR	0.30	23.04
Coherence + Interpretability + Exactness	BRM	0.34	8.61
	DSRMM	0.33	12.48
	STR	0.45	8.98
	LTR	0.30	23.99

Table 16
The performance for coherency measure (combining with the original ranking function) over hard queries in terms of *NDCG*@20.

Coherency	Ranking function	*description/ table*	*schema*	*row*	*column*	*cell*
Neural embedding (PV)	BRM	0.31	0.30	0.31	0.32	0.307
	DSRMM	0.28	0.30	0.29	0.31	0.26
	STR	0.39	0.39	0.39	0.41	0.37
	LTR	0.24	0.25	0.24	0.24	0.24
Neural embedding (BoWE)	BRM	0.31	0.30	0.31	0.33	0.30
	DSRMM	0.28	0.29	0.30	0.30	0.27
	STR	0.41	0.36	0.30	0.39	0.36
	LTR	0.26	0.23	0.24	0.23	0.23
Deep contextual (multiplication)	BRM	0.32	0.29	0.29	0.30	0.29
	DSRMM	0.28	0.31	0.26	0.26	0.25
	STR	0.42	0.40	0.38	0.38	0.37
	LTR	0.25	0.28	0.24	0.26	0.26
Deep contextual (maximum)	BRM	0.31	0.31	0.31	0.31	0.31
	DSRMM	0.29	0.29	0.29	0.29	0.29
	STR	0.41	0.41	0.41	0.41	0.41
	LTR	0.24	0.24	0.24	0.24	0.24
Statistical	BRM	0.30	0.31	0.32	0.29	0.32
	DSRMM	0.26	0.29	0.29	0.27	0.25
	STR	0.40	0.40	0.38	0.40	0.38
	LTR	0.23	0.26	0.27	0.26	0.25
Topical	BRM	0.32				
	DSRMM	0.26				
	STR	0.4				
	LTR	0.25				

Finding 6: We summarize our findings on the synergistic impact of our proposed measures as follows:

- The best variations of our proposed qualitative measures can consistently improve the performance of the baseline when interpolated with each other (except −1.25% for STR on the interpolation of coherence and interpretability).
- While the interpolation of the best variation of all three measures leads to improvement over the baselines, these improvements are not the maximum achievable results. We find the interpolation of the pair of coherence and exactness shows the overall best performance.
- The best performing interpolated models are STR and LTR based on the coherence and exactness measures. This variation shows both the best performance on *NDCG*@20 (0.50 and 0.31) and the highest percentage of improved performance over the baseline (22.94% and 26.48%).

6. Discussion

The main focus of our work is to explore how the qualitative characteristics of tables can be exploited to estimate their relevance to user queries. The motivation is that qualitative characteristics impact on whether a user finds a table appropriate for their information needs especially over a set of hard queries. We propose three classes of qualitative characteristics, namely coherency, interpretability, and exactness and formalize them from different perspectives. Coherency measures the consistency of information provided within a given table and its different modalities, interpretability refers to the degree to which the information presented in a table is understandable, and exactness considers how precisely a table can correspond to a users' information needs. Table 17.

We benefit from these qualitative characteristics to re-rank the retrieved list of tables through four state-of-the-art retrieval methods, namely, BRM, DSRMM, STR, and LTR, and empirically showed that re-ranked list of tables have a higher retrieval effectiveness over hard queries. Our assessments were performed based on the WikiTables test collection introduced in [6]. We further demonstrated that the qualitative measures were synergistic and led to even higher performance improvements over the baselines when interpolated with each other. Table 18.

In summary, our key findings include:

Table 17
The performance for each interpretability measure (combining with the original ranking function) over hard queries in terms of *NDCG*@20.

Interpretability	Ranking function	Surrounding text-based	Link-based
Neural embedding (PV)	BRM	0.31	0.30
	DSRMM	0.3	0.33
	STR	0.4	0.43
	LTR	0.30	0.27
Neural embedding (BoWE)	BRM	0.31	0.31
	DSRMM	0.29	0.29
	STR	0.41	0.41
	LTR	0.24	0.25
Deep contextual	BRM	0.31	0.31
	DSRMM	0.30	0.30
	STR	0.40	0.41
	LTR	0.25	0.25
Statistical	BRM	0.31	0.32
	DSRMM	0.34	0.34
	STR	0.39	0.40
	LTR	0.30	0.28

Table 18
The performance for each exactness measure (combining with the original ranking function) over hard queries in terms of *NDCG*@20.

Exactness	Ranking function	*description*	*schema*	*row*	*column*	*cell*
Neural embedding (BoWE) *summation*	BRM	0.32	0.30	0.31	0.32	0.31
	DSRMM	0.31	0.28	0.31	0.30	0.31
	STR	0.44	0.40	0.44	0.38	0.39
	LTR	0.26	0.26	0.24	0.23	0.25
Neural embedding (BoWE) *maximum*	BRM	0.31	0.31	0.32	0.32	0.32
	DSRMM	0.30	0.32	0.31	0.33	0.33
	STR	0.40	0.44	0.45	0.43	0.43
	LTR	0.25	0.26	0.28	0.28	0.28
Neural embedding (PV)	BRM	0.31	0.31	0.32	0.31	0.32
	DSRMM	0.31	0.35	0.32	0.32	0.33
	STR	0.39	0.44	0.44	0.40	0.44
	LTR	0.27	0.29	0.30	0.29	0.30
Deep contextual	BRM	0.31	0.31	0.32	0.31	0.31
	DSRMM	0.30	0.30	0.29	0.29	0.29
	STR	0.41	0.41	0.41	0.41	0.43
	LTR	0.24	0.24	0.25	0.25	0.24
Statistical	BRM	0.31	0.31	0.35	0.31	0.34
	DSRMM	0.26	0.28	0.35	0.29	0.36
	STR	0.38	0.45	0.46	0.43	0.48
	LTR	0.25	0.27	0.30	0.30	0.31

1. The application of qualitative measures lead to improvement of the retrieval of hard queries in ad hoc table retrieval. Depending on which perspectives and which table modalities are used for qualitative measures, the amount of acquired improvements is different.
2. Interpretability measures show the least degree of impact on the performance of the baselines, especially on the BRM and STR baselines. In contrast, the exactness measures show the strongest performance improvement on all baselines.
3. The interpolation of the best variations of our proposed coherence and exactness measures leads to noticeable improvements over the best performing state-of-the-art methods, including BRM, DSRMM, STR, and LTR. The improvements are up to 22.94% on STR method with an NDCG@20 of 0.5, which is superior to the performance of any state-of-the-art baselines for hard queries in ad hoc table retrieval.

We can also note that the proposed quality characteristics most effectively help those baseline methods that do not capture a similar notion within the baseline retrieval process. For instance, the highest degree of improvements by our proposed quality characteristics are observed over the LTR baseline method. This can be explained by the fact that LTR focuses on features that are syntactical and structural by nature, like the number of rows, columns, and empty cells, and the query word frequency in specific parts of the table, like the leftmost column, or second-to-left column. This method does not take any conceptual relations between the table and query spaces into account. As such, a significant level of improvement is observed when our proposed characteristics are incorporated with LTR.

A similar observation can be made for the BRM baseline method. Within BRM, the method implicitly includes a mechanism to identify and benefit from the most relevant modality from the table in relation to the input query. As such, our proposed exactness measure does not have as strong of an impact on BRM when compared with the other baselines. In the other baselines, all table modalities are considered in the same vein and the impact of a specific modality on the input query is not considered. As such, our proposed exactness measure can noticeably improve the performance of the baselines. For instance, the DSRMM baseline method considers all the modalities to compute the relevance score, but, as mentioned before, it limits itself to the first fifty tokens from table description, the first thirty tokens from the table schema, and twenty tokens from table rows and columns. Not only does DSRMM not consider the most suitable modality when computing relevance but it may also lose some aspects of table information due to its token limitations. Therefore, our proposed exactness measure can show a high improvement over DSRMM.

We also note that while the STR method explicitly captures semantic associations between the table and query spaces, it does not explicitly consider the role of modality alignment when computing relevance. As such, our proposed quality characteristics like exactness that take modality into account can improve the performance of STR method. Overall, we find the improvements observed over each baseline is due to the additional aspects of relevance between the query and table spaces that are not captured by the baseline ranking methods.

7. Concluding remarks

In this paper, we have proposed three classes of qualitative measures for improving the performance of state-of-the-art ad hoc table retrieval methods on hard queries. We empirically show that our proposed methods can not only show improved performance over the baselines individually but also show even higher performance improvements when systematically interpolated with each other. We specifically find that the interpolation of the best variations of our proposed coherence and exactness measures leads to noticeable improvements over the best performing state-of-the-art methods, including BRM, DSRMM, STR, and LTR. The findings of our paper indicate that hard queries are associated with relevant tables that have specific characteristics that are not directly associated with a measure of relevance. These characteristics include the coherence of the content in the table, and the exactness of the information conveyed by the table. As such, measures that consider relevance as the only criterion to rank tables will not be able to effectively rank tables in response to hard queries. We show that the consideration of such characteristics will lead to improved performance in practice and over several state-of-the-art baseline methods.

CRediT authorship contribution statement

Maryam Khodabakhsh: Conceptualization, Software, Validation, Investigation, Formal analysis, Writing - original draft. **Ebrahim Bagheri:** Conceptualization, Methodology, Investigation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Chapman, E. Simperl, L. Koesten, G. Konstantinidis, L.-D. Ibáñez, E. Kacprzak, P. Groth, Dataset search: a survey, The VLDB Journal 29 (1) (2020) 251–272.

[2] S. Zhang, K. Balog, Web table extraction, retrieval, and augmentation: A survey, ACM Transactions on Intelligent Systems and Technology (TIST) 11 (2) (2020) 1–35.
[3] M.J. Cafarella, A. Halevy, D.Z. Wang, E. Wu, Y. Zhang, Webtables: exploring the power of tables on the web, Proceedings of the VLDB Endowment 1 (1) (2008) 538–549.
[4] E. Bagheri, F. Al-Obeidat, A latent model for ad hoc table retrieval, Advances in Information Retrieval (2020) 86–93.
[5] L. Deng, S. Zhang, K. Balog, Table 2vec: Neural word and entity embeddings for table population and retrieval, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 1029–1032.
[6] S. Zhang, K. Balog, Ad hoc table retrieval using semantic similarity, in: Proceedings of the 2018 World Wide Web Conference, 2018, pp. 1553–1562.
[7] C.S. Bhagavatula, T. Noraset, D. Downey, Methods for exploring and mining tables on wikipedia, in: Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics, 2013, pp. 18–26.
[8] M. Günther, M. Thiele, J. Gonsior, and W. Lehner, "Pre-trained web table embeddings for table discovery," Fourth Workshop in Exploiting AI Techniques for Data Management, pp. 24–31, 2021.
[9] M. Khodabakhsh, E. Bagheri, Semantics-enabled query performance prediction for ad hoc table retrieval, Information Processing & Management 58 (1) (2021) 102399.
[10] J.M. Ponte, W.B. Croft, A language modeling approach to information retrieval, in: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998, pp. 275–281.
[11] S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, M. Gatford, et al., "Okapi at trec-3," Nist Special Publication Sp, vol. 109, p. 109, 1995.
[12] M. Trabelsi, Z. Chen, B.D. Davison, J. Heflin, A hybrid deep model for learning to rank data tables, in: 2020 IEEE International Conference on Big Data (Big Data), 2020, pp. 979–986.
[13] Z. Chen, M. Trabelsi, J. Heflin, Y. Xu, B.D. Davison, Table search using a deep contextualized language model, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 589–598.
[14] R. Shraga, H. Roitman, G. Feigenblat, M. Cannim, Web table retrieval using multimodal deep learning, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 1399–1408.
[15] Y. Liu, K. Bai, P. Mitra, C.L. Giles, Tablerank: A ranking algorithm for table search and retrieval, Proceedings of the National Conference on Artificial Intelligence 22 (1) (2007) 317–322.
[16] R. Shraga, H. Roitman, G. Feigenblat, M. Canim, Ad hoc table retrieval using intrinsic and extrinsic similarities, in: Proceedings of The Web Conference 2020, 2020, pp. 2479–2485.
[17] R. Shraga, H. Roitman, G. Feigenblat, B. Weiner, Projection-based relevance model for table retrieval, in: Companion Proceedings of the Web Conference 2020, 2020, pp. 28–29.
[18] S. Zhang, K. Balog, Semantic table retrieval using keyword and table queries, ACM Transactions on the Web (TWEB) 15 (3) (2021) 1–33.
[19] Z. Li, J. Tang, T. Mei, Deep collaborative embedding for social image understanding, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (9) (2019) 2070–2083.
[20] X. Wang, P. Hu, L. Zhen, D. Peng, Drsl: Deep relational similarity learning for cross-modal retrieval, Information Sciences 546 (2021) 298–311.
[21] F. Wang, K. Sun, M. Chen, J. Pujara, P. Szekely, Retrieving complex tables with multi-granular graph representation learning, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 11–15.
[22] T.A. Nakamura, P.H. Calais, D. de Castro Reis, A.P. Lemos, An anatomy for neural search engines, Information Sciences 480 (2019) 339–353.
[23] E. Bagheri, F. Ensan, F. Al-Obeidat, Neural word and entity embeddings for ad hoc retrieval, Information Processing & Management 54 (4) (2018) 657–673.
[24] A. Godbole, D. Kavarthapu, R. Das, Z. Gong, A. Singhal, H. Zamani, M. Yu, T. Gao, X. Guo, M. Zaheer, et al., "Multi-step entity-centric information retrieval for multi-hop question answering," arXiv preprint arXiv:1909.07598, 2019.
[25] R. Sankepally, T. Chen, B. Van Durme, D.W. Oard, A test collection for coreferent mention retrieval, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018, pp. 1209–1212.
[26] F. Ensan, W. Du, Ad hoc retrieval via entity linking and semantic similarity, Knowledge and Information Systems 58 (3) (2019) 551–583.
[27] N. Arabzadeh, F. Zarrinkalam, J. Jovanovic, F. Al-Obeidat, E. Bagheri, Neural embedding-based specificity metrics for pre-retrieval query performance prediction, Information Processing & Management 57 (4) (2020) 102248.
[28] H. Hashemi, H. Zamani, W.B. Croft, Performance prediction for non-factoid question answering, in: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, 2019, pp. 55–58.
[29] N. Arabzadeh, M. Khodabakhsh, E. Bagheri, Bert-qpp: Contextualized pre-trained transformers for query performance prediction, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 2857–2861.
[30] N. Arabzadeh, X. Yan, C.L.A. Clarke, Predicting Efficiency/Effectiveness Trade-Offs for Dense vs. Sparse Retrieval Strategy Selection, Association for Computing Machinery, New York, NY, USA, 2021, pp. 2862–2866.
[31] Q. Zheng, X. Ren, Y. Liu, W. Qin, Abstraction and association: Cross-modal retrieval based on consistency between semantic structures, Mathematical Problems in Engineering 2020 (2020).
[32] X. Tu, J.X. Huang, J. Luo, T. He, Exploiting semantic coherence features for information retrieval, in: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, 2016, pp. 837–840.
[33] L. Wang, S. Li, Y. Lü, H. Wang, Learning to rank semantic coherence for topic segmentation, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1340–1344.
[34] J. Lee, J.-K. Min, A. Oh, C.-W. Chung, Effective ranking and search techniques for web resources considering semantic relationships, Information Processing & Management 50 (1) (2014) 132–155.
[35] H. Zamani, W.B. Croft, Embedding-based query language models, in: Proceedings of the 2016 ACM international conference on the theory of information retrieval, 2016, pp. 147–156.
[36] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International conference on machine learning, 2014, pp. 1188–1196.
[37] R. Zhang, J. Guo, Y. Lan, J. Xu, and X. Cheng, "Aggregating neural word embeddings for document representation," European Conference on Information Retrieval, pp. 303–315, 2018.
[38] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, vol. 1, pp. 4171–4186, 2019.
[39] K. Braunschweig, M. Thiele, E. Koci, and W. Lehner, "Putting web tables into context," Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016), vol. 1, pp. 158–165, 2016.
[40] M. Kozlowski, H. Rybinski, Clustering of semantically enriched short texts, Journal of Intelligent Information Systems 53 (1) (2019) 69–92.
[41] P. Li, T. Li, S. Zhang, Y. Li, Y. Tang, Y. Jiang, A semi-explicit short text retrieval method combining wikipedia features, Engineering Applications of Artificial Intelligence 94 (2020) 103809.
[42] M. Esposito, E. Damiano, A. Minutolo, G. De Pietro, H. Fujita, Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering, Information Sciences 514 (2020) 88–105.
[43] H.K. Azad, A. Deepak, A new approach for query expansion using wikipedia and wordnet, Information sciences 492 (2019) 147–163.
[44] Z. Wu, H. Zhu, G. Li, Z. Cui, H. Huang, J. Li, E. Chen, G. Xu, An efficient wikipedia semantic matching approach to text document classification, Information Sciences 393 (2017) 15–28.
[45] J. Hu, L. Fang, Y. Cao, H.-J. Zeng, H. Li, Q. Yang, Z. Chen, Enhancing text clustering by leveraging wikipedia semantics, in: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, 2008, pp. 179–186.

[46] X. Hu, X. Zhang, C. Lu, E.K. Park, X. Zhou, Exploiting wikipedia as external knowledge for document clustering, in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009, pp. 389–396.
[47] X. Pan, K. Sun, D. Yu, J. Chen, H. Ji, C. Cardie, and D. Yu, "Improving question answering with external knowledge," arXiv preprint arXiv:1902.00993, 2019.
[48] J. Wu, J. Lu, A. Sabharwal, and R. Mottaghi, "Multi-modal answer validation for knowledge-based vqa," arXiv preprint arXiv:2103.12248, 2021.
[49] A. Spink, D. Wolfram, M.B. Jansen, T. Saracevic, Searching the web: The public and their queries, Journal of the American society for information science and technology 52 (3) (2001) 226–234.
[50] Q. Zhou, N. Yang, F. Wei, S. Huang, M. Zhou, and T. Zhao, "A joint sentence scoring and selection framework for neural extractive document summarization," IEEE/ACM Trans. Audio, Speech and Lang. Proc., vol. 28, p. 671–681, Jan 2020.