Datenbank Spektrum Special Issue on Trends in Information Retrieval Evaluation manuscript No.
(will be inserted by the editor)

Humans optional? Automatic Large-Scale Test Collections for Entity, Passage, and Entity-Passage Retrieval

Laura Dietz · Jeff Dalton

Received: date / Accepted: date

Abstract Manually creating test collections is a time-, effort-, and cost-intensive process. This paper describes a fully automatic alternative for deriving large-scale test collections, where no human assessments are needed. The empirical experiments confirm that automatic test collection and manual assessments agree on the best performing systems. The collection includes relevance judgments for both text passages and knowledge base entities. Since test collections with relevance data for both entity and text passages are rare, this approach provides a cost-efficient way for training and evaluating ad hoc passage retrieval, entity retrieval, and entity-aware text retrieval methods.

Keywords automatic evaluation · entity and passage retrieval · complex answer retrieval

1 Introduction

Passage and entity retrieval are central components in search engine result pages (SERP), "fetch and browse" interfaces, and composite retrieval (Arvola et al., 2010; Kaszkiel and Zobel, 1997; Bota et al., 2014). However, creating reusable test collections is difficult and time consuming (Wade and Allan, 2005). The availability of large-scale entity linking methods have led to a large number of hybrid approaches that combine information about text and (Wikipedia) entities, such as people, proteins, or events. Most of these hybrid approaches combine different components for predicting the relevance of entities and text, which are integrated to improve performance on both tasks which are defined as follows:

Passage and entity retrieval tasks. Given an information need expressed as a search query Q, retrieve a ranking of (1) passages from a given corpus and (2) entities from a given knowledge graph that is ordered by relevance for the query.

Examples of hybrid entity-passage retrieval approaches either improve on passage retrieval by exploiting information about relevant entities (Dalton et al., 2014; Xiong and Callan, 2015; Xiong et al., 2017b,a; Raviv et al., 2016, *inter alia*); or improve entity retrieval by incorporating knowledge of relevant passages (Bast et al., 2018, 2016; Boston et al., 2014; Schuhmacher et al., 2015; Dietz, 2019, *inter alia*).

In the context of search result diversification, as an additional requirement, the ranking needs to cover passages (or entities) that are relevant for different potential query interpretations. In other settings, ranked results need to be clustered into coherent sub-topics. Both settings require relevance annotations for different query facets, i.e., a ground truth of relevance for each query interpretation or sub-topic.

A variation on entity-aware retrieval is the task of entity support-passage retrieval, where given a query and an entity, a ranking of passages is to be predicted that explain why the entity is relevant for the query (Blanco and Zaragoza, 2010; Chatterjee and Dietz, 2019; Arnold et al., 2019).

Closely related is entity-centric question answering where the query is a question for which the answer is

Laura Dietz
University of New Hampshire, USA
E-mail: dietz@cs.unh.edu

Jeff Dalton
University of Glasgow, UK
E-mail: jeff.dalton@glasgow.ac.uk

a relevant entity that needs to be extracted from relevant passages (Yang et al., 2015; Sawant et al., 2019). The task of retrieving such answer-containing passages is called answer-passage retrieval (O'Connor, 1980). In conversational search (Choi et al., 2018; Dalton et al., 2019), multiple questions evolve around a changing subject that is expressed through entities and passages.

While all above mentioned approaches exploit relevant entities and text, they are usually evaluated and trained on benchmarks that are designed for either ad hoc document retrieval (e.g., Clueweb or Robust04) or entity retrieval (e.g., INEX or TREC Entity), but not both. This poses non-ideal circumstances for evaluating the quality of methods and studying potential mistakes made by underlying components. It also leads to long training times due to marginalization over latent parameters (Xiong and Callan, 2015) or an explosion of the feature space (Dalton et al., 2014).

While test collections with relevance assessments for both text and entities would be beneficial for research on integrated entity-passage retrieval models, such test collections are expensive to create manually and hence not widely available (cf. Section 2.1). To study the retrieval task in the context of search result clustering or diversification, relevance assessments for query facets need to be available.

In this work, we attempt the daring experiment to build a test collection for entity-passage retrieval that does not require any human assessor. We describe a mechanism for creating fully automatic test collections for passage, entity, and integrated entity-passage ranking with query facets. If successful, our approach offers low-cost access to large-scale test collections for ranking text and entities—which is particularly important for academic research. Of course, we envision automatic test collections to be complemented with human-assessed test collections for the purpose of training, evaluation, and error analysis. To assess whether this "humans optional" approach is viable, we compare the ranking of systems (i.e,. leaderboard) produced with the automatic test collection to a gold standard leaderboard that was produced with an established approach using manual pool-based assessments.

For the experimental study, systems and manual assessments were taken from the TREC Complex Answer Retrieval track (TREC CAR, Dietz et al. 2017, 2018). The manual test collection was constructed by exhaustively assessing all top ranked documents and entities of all participating systems. The assessment procedure was performed under the supervision of experts from the National Institute for Standards and Technology (NIST) who have several decades of experience in creating manual test collections for the Text Retrieval Conference (TREC). The manual benchmark hence offers a reliable gold standard for the systems' quality.

As an example domain, we focus on fact-oriented popular science queries where users ask for overviews of multi-faceted topics. We interpret the relevance of passages and entities as "Is this passage or entity to be included in an article about the topic?" A test collection for such information needs—with query facets—is derived from a corpus of Wikipedia pages and science textbook chapters. The textbook chapters were taken from a textbook question answering corpus (TQA, Kembhavi et al. 2017). Our approach is more generally applicable beyond this concrete example, by varying the corpora of input pages such as other Wikis, other textbook chapters, product descriptions, knowledge compendia, taxonomies, or glossaries. Our approach is related to an established test collection approach in community question answering (CQA) (Shah and Pomerantz, 2010); test collections are derived from web sites like stack overflow[1] or yahoo[2] to use questions as queries and confirmed answers as true text passages. Similar benchmarks can be derived from chatbot dialog collections (Choi et al., 2018).

In general, our approach is applicable to any collection of human-authored pages that coincides with anticipated queries and responses. An example where our approach is not applicable are news articles, because news titles are not a good representation of queries (Soboroff et al., 2018).

Contributions. To advance research in the area of evaluation methods, we propose a fully automatic approach for building large-scale test collections with several hundred thousand queries. For queries Q, the test collection provides relevance data for text passages and knowledge base entities. Our test collection generation paradigm can be applied to a wide range of sources, such as Wikipedia, Web crawls, or QA sites. Unlike many other approaches towards automatic test collections, our approach does not need any human assessments and uses realistic informational queries.

An important contribution of this work is the experimental demonstration that human assessors and automatic approach agree about the relative performance of a diverse set of systems. Specifically, for the goal of ranking systems by quality (i.e., a leaderboard), the leaderboards produced with the manual benchmarks and our automatic benchmark agree nearly perfectly, achieving Kendall's τ of 0.93 for text passage retrieval and 0.89 for entity retrieval.

[1] `https://archive.org/details/stackexchange`
[2] `https://webscope.sandbox.yahoo.com/catalog.php?datatype=l`

Outline. Section 2 elaborates on existing test collections and related approaches for benchmark creation. Section 3 introduces our approach for automatic test collection creation and a discussion of different error modes. Experiments in Section 4 evaluate the quality of the automatic benchmark with human assessors on passages and entities, before concluding in Section 5.

2 Related Work

We survey related test collections for combined document and entity retrieval. Then we discuss approaches to reduce the number of manual assessments as well as fully automatic pseudo-test collection approaches.

2.1 Manual Entity and Passage Test Collections

The TREC Web Track provides benchmarks, assessed on deep pools and a large corpus. A collection of entity link annotations was released (FACC1, Gabrilovich et al. 2013), although without any relevance data.

Test collections for entity retrieval are from TREC Entity (Balog and Neumayer, 2013) and INEX (Demartini et al., 2009) tracks. Some focus on text retrieval for queries about people or person entities, such as TREC Knowledge Base Acceleration (Frank et al., 2014). Bast et al. (2018) evaluate a “KB+text” system via entity ranking benchmarks. Only very few test collections provide data for both text relevance and entity relevance. Only small add-on test collections are available TREC Web and Robust04 topics (Schuhmacher et al., 2015; Foley et al., 2016). In this work, we develop an approach to build large and reliable test collections to evaluate approaches that combine entity and text retrieval.

Passage retrieval has been studied within a range of data sets, such as the TIPSTER (Callan, 1994), TREC HARD track (Allan, 2003), and INEX Focused Task (Kamps et al., 2008) and “Relevant in Context” (Kamps et al., 2007). Such datasets do not pre-define passage boundaries, and hence use character-based evaluation metrics to reuse existing passage annotations (Wade and Allan, 2005). While such measures can also be applied to the automatic test collections created by our method, our work focuses on how to create a test collection independently of customized evaluation measures.

2.2 Automatic Support for Manual Test Collections

Several approaches for reducing the assessment costs have been explored. A system of Cormack et al. (1998) aids manual assessors in determining the relevance through interactive searching and judging. Jayasinghe et al. (2014) suggest a machine learning method to obtain more resilient assessment pools for manual assessment. Yilmaz et al. (2008) reduce manual assessment costs by sampling assessment pools randomly from input rankings while preferring highly ranked documents. They suggest extensions to MAP and nDCG that correct sampling bias and obtain better performance estimates than random sampling.

Given a small number of manual assessments for a query, the AutoTAR algorithm (Zhang et al., 2018) trains a query-specific classifier. The classifier is used to identify documents with similar characteristics, which are presented to the assessor for assessment. Machine learning, pool prediction, and manual assessment are interleaved in a continuous process.

While these approaches reduce the number of required manual assessments, benchmark creation still hinges on human assessors. In contrast, our automatic test collection approach does not require any human intervention beyond the selection of suitable input pages.

2.3 Fully Automatic Pseudo-Test Collections

The closest in spirit to our work are approaches towards fully automatic test collections, which are also called pseudo-test collections. In general, approaches are based on selecting subsets of a corpus that represent relevant documents for possible information needs, from which queries are derived. Two main approaches for query derivation are to simulate queries from term distributions and to exploit available meta-annotations.

Queries are simulated by selecting terms that maximize the probability of discriminating between the relevant and non-relevant document set (Berendsen et al., 2013; Azzopardi et al., 2007). Unfortunately, there is no guarantee that such queries are realistic.

An alternative is to derive the query from meta-annotations of the corpus. A wide range of meta-annotations have been explored, such as anchor text (Asadi et al., 2011), metadata of scientific articles about method, classification, and control (Berendsen et al., 2012), categories in the Open Directory Project (Beitzel et al., 2003), or glosses in Freebase (Dalvi et al., 2015).

While automatic test collections can support both the evaluation and training of approaches, most related work focuses on one or the other. Berendsen et al. (2012) report a discrepancy between test collection quality for evaluation versus training purposes.

Our work derives queries from meta-annotations in the form of titles and headings. In contrast to prior work, our experimental study is integrated into a shared

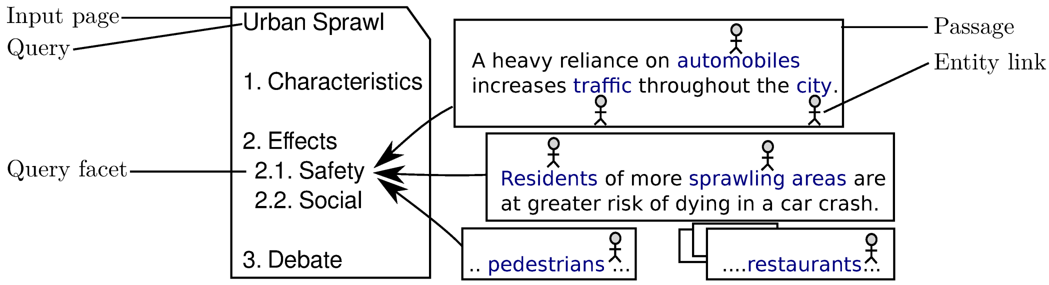


Fig. 1: Example for test collection creation taken from the Wikipedia page about Urban Sprawl. Left: Page title and outline from which query and query facets are derived. Right: Text passages with entity links, where arrows depict the relevant passages. Linked entities in relevant passages are automatically marked as relevant.

evaluation at TREC: Automatically derived queries are also used in the manual assessment. For training and evaluation, participants had access to both automatic and manual benchmarks (evaluation benchmarks were only released after the evaluation). This allows for an in-depth comparison between manual and automatic benchmarks for both passage and entity ranking tasks.

3 Automatic Creation of Test Collection for Queries and Facets

Our fully-automatic approach creates a test collection for evaluating passage ranking, entity ranking, and integrated entity-passage ranking—in some cases even with query facets. Our approach relies on a collection of a human-created corpus, which is often readily available, such as a Wikipedia dump, textbook chapters, product descriptions, a knowledge compendium, or glossary.

3.1 Test Collection Format

The test collection consists of a passage corpus, a reference knowledge base, and relevance data in the form of tuples that contain:

- query text and ID,
- passage ID and/or entity ID,
- binary relevance (relevant vs. non-relevant), and
- query facet for this assessment (optional)

Relevance data for established evaluation frameworks (such as “qrels” for trec_eval[3]) or learning-to-rank tools such as RankLib[4] or SVMrank[5] can be derived from this representation. While useful for method development of integrated entity-passage approaches, to evaluate the passage quality, the entity information is ignored—and vice versa for entity ranking evaluation. For entity support-passage retrieval, query and entity are given, and the resulting ranking is evaluated based on the passage and relevance information. For result diversification, query facets are used in intent-aware evaluation measures; for sub-topic clustering, query facets provide ground truth information for predicted clusters.

[3] https://github.com/usnistgov/trec_eval
[4] http://www.lemurproject.org/ranklib.php
[5] https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

3.2 Derivation from Input Sources

Given a manually created source corpus of input pages, we suggest the following automatic approach for deriving a test collection. An example of an input page about the topic “Urban Sprawl” is depicted in Figure 1. A summary of the approach is given in Algorithm 1.

Queries are derived from the titles of input pages, which we refer to as $title in the following. In our example domain, the information need is interpreted as “Provide comprehensive information about $title”. Hence, the content of a Wikipedia page on Urban Sprawl is relevant for this information need. After identifying headings, the remaining content of the page is split into paragraph-sized passages with content-based ID. The passage corpus is comprised of all passages from a large collection of input pages, which also include pages for queries. Only passages contained in the query-generating page are defined as relevant for the query.

Entities that are mentioned on this page are assumed to be relevant to a user that would like to know more about the query. For the example query Urban Sprawl, relevant entities are automobiles, city, traffic, resident, sprawling areas, car crash, restaurants, etc. In the case of Wikipedia, such relevant entities will be indicated as a hyperlink to the entity’s Wikipedia page. For input pages from other sources, an entity linking tool (such as Tagme, Ferragina and Scaiella 2010 or DBpedia Spotlight, Mendes et al. 2011) is used to identify entities that are mentioned on the input page—which we define as relevant entities. To define the universe of possible entities, the test collection provides a reference knowledge base. A common choice are knowledge bases derived from Wikipedia or the Wikipedia corpus itself, which is applicable to both relevance criteria, as long as appropriate Wikipedia dump versions are used.

By keeping track of passages in which entities are mentioned, we align relevant passages with relevant entities which is useful to study the context of relevant entities as well as entity-passage ranking tasks. In a similar fashion, benchmarks for many related tasks can be derived such as text segmentation across section boundaries (Arnold et al., 2019) or query-centric passage clustering (Wan and Yang, 2008).

When input pages have sections with headings, we use these headings to define query facets. In hierarchical sections, we either define the lowest level of sections as facets (as in TREC CAR year 1) or any hierarchical level (as in TREC CAR year 2). In either case, the heading of a section is concatenated with parent headings and page title to define the query facet. In the example of Urban Sprawl (Figure 1), five query facets are defined: “Urban Sprawl/ Characteristics”, “Urban Sprawl/ Effects/ Safety”, “Urban Sprawl/ Effects/ Social”, “Urban Sprawl/ Effects”, and “Urban Sprawl/ Debate”. The relevance assessment of passages and entities in the content of the corresponding section is annotated with the query facet. Query facet annotations allow to evaluate the coverage of different query facets in rankings about the information need through intent-aware evaluation measures (Sakai and Song, 2011). Alternatively, specific queries can be derived by concatenating page title and headings, such as “Urban Sprawl Effects Safety”, for which only passages and entities in the corresponding section (or a sub-section) are marked as relevant. The latter approach was taken to derive queries in TREC CAR year 1 and year 2.

If Wikipedia is used to derive both queries and the reference knowledge base, there is a danger of releasing the true answers to queries as part of the reference knowledge base. To avoid inadvertent cheating in the evaluation, input pages from which queries are derived, must be removed from the reference knowledge base. An alternative is to use a different collection of input pages, such as school books from the TQA dataset and reserve the use of Wikipedia as a reference knowledge base only. The former approach was applied in TREC CAR year 1 and the “Wiki-18” subset of year 2. The latter approach was applied in the “TQA” subset of TREC CAR year 2.

3.3 Discussion

While the automatic test collection is intended to be used for relative system comparisons, the absolute value of an evaluation measure is expected to be lower than in manual collections. This is due to a set of potentially relevant passages which might not have been included on the input page by chance, and thus are counted as non-relevant by our construction. This is in contrast to manual test collections, where assessors will judge the relevance of every passage in the assessment pool. If there are several similar passages, under a manual assessment all would be counted as relevant, where under the automatic collection only the one included on the input page would be counted as relevant—we call the remaining passages *false non-relevant assessments*, since they are false negatives when the manual benchmark is taken as the gold standard.

To reduce the problem of false non-relevant assessments, it is essential to detect near-duplicates in the passage corpus and treat the whole set of duplicates as relevant whenever one of its members is relevant according to our criterion. Despite deduplication efforts, the evaluation in Section 4 reveals that human judges are more lenient than the automatic approach, yielding about three times as many positive relevance assessments per query than the automatic approach. In the following, we discuss three arguments why this issue does not negatively affect relative system evaluation through the automatic test collection.

The creation of manual assessments is always a noisy process since some assessors are strict while others are lenient; furthermore, changes in concentration and expertise will impact the resulting assessments. Moreover, whenever a relevant passage is not contained in the assessment pool, such passages will be assumed to be non-relevant. Hence, false non-relevant assessments arise in manual evaluation as well, albeit due to different effects. Consequently, the absolute value of an evaluation measurement needs to be taken with a grain of salt: Only relative performance improvements over a baseline are reliable indicators of performance. In this context, the automatic test collection behaves similarly to a strict assessor with an imperfect assessment pool.

One might be worried that a low number of positive assessments per query would lead to unstable performance evaluations. However, this low number is compensated by a larger number of test queries (hundreds to millions) in comparison to manual assessments (fifty to a few hundred in public test collections). As a result, unfair penalization of systems that retrieve false non-relevant assessments applies randomly across all systems. With the help of statistical analyses, stable rel-

Algorithm 1 Automatic test collection creation.

Given set of query pages, derive the automatic test collection:

- **Passage corpus:** All passages from all pages after redundancy removal, in randomized order.
- **Knowledge base:** All Wikipedia pages, except query pages.
- **Queries:** Titles of query pages, e.g., "Urban Sprawl"
- **Passage relevance:** All passages on the page are defined as relevant for the query (page-level passage retrieval).
- **Entity relevance:** All entities to which the page contains an entity link are defined as relevant (page-level retrieval).

In the case where input pages have sections with headings, query facets are derived from the section hierarchy.

- **Query facets:** Each section heading in the outline of the page, e.g., "Urban Sprawl/ Effects/ Safety".
- **Faceted passage relevance:** All passages located in the corresponding section (or sub-section) defined as relevant with respect to the facet (section-level passage retrieval).
- **Faceted entity relevance:** Entities to which the corresponding section contains an entity link are defined as relevant for the facet (section-level entity retrieval).

In the example of TREC CAR year 1 and year 2, the queries are formed from query facets as "$title/$heading". Furthermore, a test collection for entity support-passages is derived from the information above by considering each pair of query facet and relevant entity as a new information need for which the passages that mention the corresponding entity on the input page are defined as relevant.

ative performance estimates can be derived as long as the number of overall assessments is on the same order of magnitude (cf. total positive relevance assessments in Tables 1 and 2).

An important issue in information retrieval is the assessment of marginal relevance, where redundant passages should not be regarded as relevant. The automatically created benchmark is naturally suitable for evaluating marginal relevance, since a human-edited input page is unlikely to contain many redundancies.

3.4 Implementation Details

To reduce the issue of false non-relevant assessments, near-duplicates are removed from the passage corpus. We identify sets of redundant passages with a combination of GloVE-based (Pennington et al., 2014) locality-sensitive hashing and a 50% bigram-overlap criterion. We replace all occurrences of near-duplicate passages with a representative, which is reflected in the passage corpus and relevance data.

We manually select input pages to derive queries from, to ensure that page titles and headings represent realistic search queries in both Wikipedia and textbooks from the TQA corpus. We omitted pages with quality issues or less than three sections and reformulated headings where necessary.

4 Experimental Evaluation

We study to which extent such an automatic test collection can substitute manual assessments for the purposes of evaluating passage retrieval and/or entity retrieval methods. While our framework can derive relevance data for several different tasks, here we evaluate the query-faceted benchmark in the context of the TREC Complex Answer Retrieval track (CAR). We compare the leaderboards under automatic and manual test collection with respect to number of systems swapped (Kendall's τ) and changes in the rank of a system (Spearman's rank correlation coefficient ρ). Furthermore, we use Cronbach's $\hat{\alpha}$ to analyze the reliability of both test collections Bodoff (2008).
We study the following research questions:

- RQ1: Does the automatic passage test collection yield the same leaderboard of systems as a manual test collection?
- RQ2: What are the effects of false non-relevant assessments (as discussed in Section 3.3)?
- RQ3: Does the automatic entity test collection yield the same leaderboard as a manual test collection?
- RQ4: What is the effect of creating entity test collections through an entity linking tool versus manually edited hyperlinks?

We discuss the dataset and evaluation paradigm below and elaborate on results for each research question.

4.1 Example Data Set: TREC CAR

In TREC CAR year 1 and year 2, the track provides queries in the form "$title/$heading" to participants, which coincide with query facets as defined in our approach. (The "$title" queries were used in year 3.) Using the same set of queries, the track offers both an ad hoc passage retrieval task and an ad hoc entity retrieval task. This evaluation uses submitted systems and pool-based assessments from the passage task of year 1 (Dietz et al., 2017) and entity task of year 2 (Dietz et al., 2018). In both cases, manual assessments were created by NIST assessors.

A Wikipedia dump from Dec 20, 2016 (Wiki-16) is used as input pages to create the passage corpus and the reference knowledge base which defines the legal set of entities. Pages from which queries are derived are omitted from the reference knowledge base.

In year 1, query facets are selected from Wiki-16 input pages, where in year 2, query facets are selected from textbook chapters of the TQA dataset[6] and new

[6] `http://data.allenai.org/tqa/`

Table 1: Passage retrieval benchmark statistics of training data (benchmarkY1train) and test collection (benchmarkY1test). Manual assessments are produced with submitted systems in TREC CAR year 1. The inter-annotator agreement κ is comparable to other IR experiments (Alonso and Mizzaro, 2012).

	Passage Train Data	Passage ⋆
Size of passage corpus, duplicates removed	⟵ 29,678,367 ⟶	
Automatically assessed queries ($title)	117	133
Automatically assessed query facets ($title/$heading)	1,816	2,125
Total positive automatic relevance assessments	4,530	5,820
Manually assessed queries ($title)	–	132
Manually assessed query facets ($title/$heading)	–	702
Total positive manual assessments (must, should, can)	–	7,796
Total negative manual assessments (topic, non-relevant, trash)	–	23,389
Binary inter-annotator agreement (Dietz et al., 2017)	–	Fleiss $\kappa = 0.57$

Table 2: Entity retrieval benchmark statistics of training data (benchmarkY1train) and both subsets of the test collection (benchmarkY2test). Manual assessments are produced with submitted systems in TREC CAR year 2.

	Entity Train Data	TQA Entity ⋆	Wiki-18 Entity ⋆
Size of knowledge base, omitting input pages for queries	⟵ 5,153,990 ⟶		
Automatically assessed queries ($title)	117	31	34
Automatically assessed query facets ($title/$heading)	1,816	277	699
Total positive automatic relevance assessments	13,031	1,727	15,317
Manually assessed queries ($title)	–	18	9
Manually assessed query facets ($title/$heading)	–	128	143
Total positive manual assessments	–	1,817	1,356
Total negative manual assessments	–	1,858	3,384
Binary inter-annotator agreement (Dietz et al., 2018)	–	Fleiss $\kappa = 0.42$ (without annotator 2)	

pages from a 2018 Wikipedia dump (Wiki-18). The change was necessary after the reference knowledge base was released containing all Wiki-16 pages except those used for the year 1 benchmark.

While the TREC CAR track features several benchmarks, we focus on these two datasets since both a manually assessed benchmark and an automatic test collection is available for their query set. In contrast, no manual entity benchmark was created for year 1. Furthermore, automatic passage relevance assessments cannot be derived from input pages used in year 2, as query pages were derived from pages in Wiki-18 that did not exist in Wiki-16, while the passage corpus was created from Wiki-16.

Systems used in our experimental evaluation were submitted by participants to the TREC CAR track and include neural ranking methods, entity-aware ranking methods, standard retrieval models such as BM25, RM3, and sequential dependence models, as well as methods based on learning-to-rank. Participants were prohibited to access a Wikipedia corpus except the provided training data and reference knowledge base—neither of which included the input pages from which queries were derived. We anonymize[7] the systems since they are not the focus of this work, details are available in the TREC proceedings and TREC CAR overview notebooks (Dietz et al., 2017, 2018). System runs are provided upon request.

The manual assessment was conducted on assessment-pools of the top six passages and top five entities from each contributed system. Passages that are relevant under the automatic benchmark were added to the manual assessment pool for verification. Some rankings contained fewer passages and several passages were in-

[7] Passage task: `A--1`: mpii-nn4_pos_hperc, `A--2`: mpii-nn6_pos, `A--3`: mpii-nn6_pos_tprob, `B--1`: CUISPR, `C--1`: UNH-benchmarkY1test.expan, `C--2`: UNH-benchmarkY1test.bm25, `D--1`: UTDHLTRINN20, `D--2`: UTDHLTRINN50, `D--3`: UTDHLTRIAR, `E--1`: treccarict, `F--1`: nyudl-qr, `F--2`: nyudl-ds, `F--3`: nyudl-qrds, `G--1`: ECNU-runONE.
Entity task: `A--1`: UNH-e-L2R, `A--2`: UNH-e-graph, `A--3`: UNH-e-mixed, `B--1`: uog-paragraph-rf-ent, `B--2`: uog-linear-ltr-hier-ent, `B--3`: uog-heading-rh-sdm-ent, `C--1`: DWS-UMA-EntAspQLrm, `C--2`: DWS-UMA-EntAspBM25none, `D--1`: CUIS-dogeDodge, `D--2`: CUIS-XTS, `D--3`: CUIS-Swift.

cluded in multiple rankings. On average, 44 passages and 31 entities per query facet were assessed.

Given the query and facet, assessors judge whether a passage or entity is to be included in an article about the topic. The grading scale differentiates between must be included (3), should be included (2), can be included (1), roughly on topic but not sufficiently specific (0), not relevant (-1), trash (-2). In this evaluation, we only distinguish positive assessments (must, should, can) from negative assessments (topic, not-relevant, trash).[8]

4.2 Evaluation Paradigm

The motivation for our automatic test collection approach is to evaluate systems without human involvement. We experimentally evaluate whether the automatic test collection and manual assessors agree on the relative quality of systems. We use both automatic and manual relevance data to evaluate systems submitted by track participants and predict the leaderboard, i.e., ranking of systems by relative performance. In line with the TREC CAR guidelines, system performance is evaluated with R-Precision (RPrec), Mean-average precision (MAP), and Normalized Discounted Cumulative Gain (nDCG@20) as implemented in the evaluation tool trec_eval (with '-c' option). TREC participants could train their submitted systems on the automatic benchmark of dedicated training queries called benchmarkY1train which includes training data for both passage and entity ranking tasks. Statistics on training data is presented in Tables 1 and 2. Benchmarks marked with $\star$ are used in this evaluation and were released only after the shared task concluded. Furthermore, a much larger automatic test collection based on 285,000 input pages is available in the TREC CAR data release to support the training of neural networks.[9]

The intuition behind Chronbach's $\hat{\alpha}$ is that the variance of system scores across queries is representing how consistent the difficulty of queries and their assessments are. Hence, smaller variances mean that assessments for each query in the collection are of similar difficulty—as opposed to having a high variance in leniency of assessors and difficulty of queries. Bodoff (2008) states that resulting $\hat{\alpha}$ only "pertains to that particular group of algorithms [systems]", hence we only use it to compare between manual and automatic test collections and use the same set of systems to calculate Cronbach's $\hat{\alpha}$.

[8] Similar results are obtained when "roughly on topic" is counted as a positive assessment (Dietz et al., 2017).

[9] `http://trec-car.cs.unh.edu/datareleases/`

4.3 RQ1: Evaluation of Passage Rankings

We first evaluate the quality of the automatic passage test collection. Figures 2a and 2d demonstrate that both automatic and manual relevance data sets result in nearly the same leaderboard.

Very high Spearman's rank correlation and Kendall's τ of 0.93 are obtained for all three evaluation measures (Table 4). According to Voorhees (2001), a τ of 0.9 suggests that the ordering of systems under both benchmarks are not meaningfully different. In fact, the only difference is that a single system fell two ranks behind.

The reliability measure Cronbach's $\hat{\alpha}$ for both automatic and manual test collection is comparable. Both are even slightly higher than in early TREC collections (cf. Table 3 in this paper with Table 2 of Bodoff (2008)).

We conclude that in this example task, automatic relevance data is as suitable for evaluating passage rankings as manual test collections produced with trained assessors.

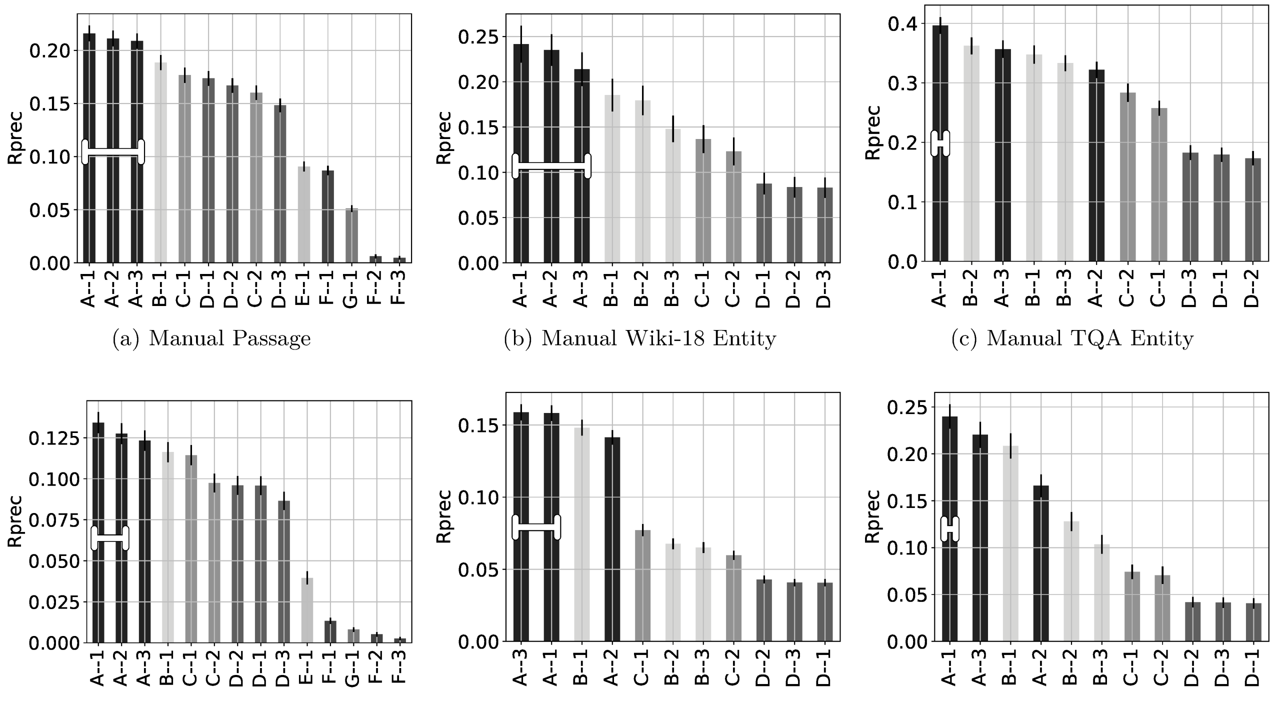
4.4 RQ2: Missing Relevant Passages Vs. Size

As discussed in Section 3.3, we study false non-relevant assessments which arise since for every page, one can easily imagine an alternative version that uses different words and selects different examples but is equally useful to the reader. This can potentially lead to a large amount of false non-relevant assessments, despite deduplication efforts.

Indeed, Figures 2a and 2d show that the evaluation scores are generally lower under the automatic relevance data. The manual pooled evaluation found three times as many relevant passages: On average each query facet has 10.7 positive manual assessments, but only 2.7 in the automatic test collection. However, nearly all automatic relevant assessments were confirmed as relevant by manual assessors (except 1%). The 1% exception are passages that were separated from their context such as "See the example below".

On the whole, automatic and manual assessments agreed on 79% of assessments; resulting in an inter-annotator agreement between automatic and manual assessments of Cohen's $\kappa = 0.268$. Nearly all disagreement is due to the higher strictness levels of the automatic benchmark, which renders the automatic benchmark more challenging with respect to absolute values of evaluation measurements. However, this strictness affected all system's scores equally and resulted in nearly the same relative order of systems on the leaderboard.

The system's performance differences are consistent across all queries, reflected in Chronbach's $\hat{\alpha}$ (Table 3). A useful test collection must be able to discriminate



(a) Manual Passage (b) Manual Wiki-18 Entity (c) Manual TQA Entity

(d) Auto Passage (e) Auto Wiki-18 Entity (f) Auto TQA Entity

Fig. 2: Leaderboard of passage and entity runs under manual (top) and automatic (bottom) test collections. The white bracket indicates systems for which no significant difference to the best system could be detected with a paired-t-test ($\alpha = 5\%$). The systems are anonymized to "team–run", all runs by the same team share the same bar color. See footnote and TREC CAR Overview reports for details about the systems.

Table 3: Reliability measure Cronbach's $\hat{\alpha}$ with respect to RPrec evaluation scores of systems. For comparison, the reliability of TREC 3-10 collections range between 0.857 and 0.933 (Bodoff, 2008).

	Passage	TQA Entity	Wiki-18 Entity
Manual test collection	0.996	0.981	0.959
Automatic test collection	0.993	0.989	0.996

Table 4: Kendall's τ and Spearman's ρ correlations between automatic and manual leaderboards for passage ranking and entity ranking tasks.

	τ_{RPrec}	τ_{MAP}	τ_{nDCG}	ρ_{RPrec}	ρ_{MAP}	ρ_{nDCG}
Passage	0.93	0.93	0.93	0.98	0.99	0.98
TQA Entity	0.74	0.85	0.89	0.89	0.93	0.96
Wiki-18 Entity	0.74	0.67	0.81	0.90	0.86	0.93

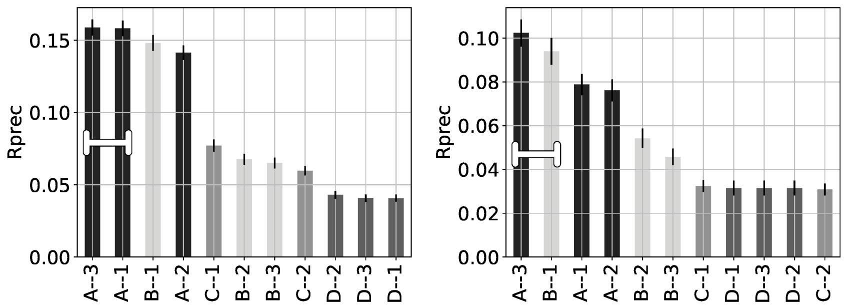
better from worse systems through significance analyses. A paired-t-test significance analysis based on the system with the highest mean performance depicted by white brackets in Figure 2) determines a small set of clear "winners". Moreover, standard error bars are of the same relative magnitude for automatic test collection and manually created assessments, indicating the suitability of IR evaluations.

These encouraging results are obtained when comparing an automatic and a manual approach that have a comparable number of positive relevance assessments across all queries (cf. Tables 1 and 2).

While the manual assessments were limited by a budget of 240 hours for the passage task (120 hours for the entity task), the automatic test collection approach is only limited by the number of input pages.

Table 5: Kendall's τ and Spearman's ρ correlations between leaderboards on the Wiki-18 subset. Three-way comparison of entity test collections: (1) entity links, (2) hyperlinks inserted by the article editor, and (3) manually assessed benchmark.

	τ_{RPrec}	τ_{MAP}	τ_{nDCG}	ρ_{RPrec}	ρ_{MAP}	ρ_{nDCG}
Using Entity Linking versus Manual Assessments	0.74	0.67	0.81	0.90	0.86	0.93
Using Editor's Hyperlinks versus Manual Assessments	0.63	0.78	0.85	0.80	0.90	0.94
Entity Linking versus Editor's Hyperlinks	0.74	0.85	0.89	0.90	0.94	0.96



(a) Entity Linking Tool (b) Editor's Hyperlinks

Fig. 3: Entity ranking leaderboard under the automatic test collection for the Wiki-18 subset. The relevance of entities is determined with an entity linking tool (left) versus Wikipedia hyperlinks created by the article editor (right).

This evaluation is based on automatic test collections derived from fewer than 150 manually selected input pages, an alternative train/test set for TREC CAR was derived from 285,000 input pages.

4.5 RQ3: Evaluation of Entity Rankings

Finally, we evaluate the quality of the automatic entity test collection. Queries in TREC CAR year 2 are derived from two different sources of input pages, Wiki-18 and TQA, which discuss similar topics, but their pages have different characteristics. TQA pages explain the topic to school children and use simplified language, where Wikipedia pages often mention specific entities and many technical details.

Figure 2 shows that for both query subsets, automatic and manual relevance data result in a very similar leaderboard. No significant difference could be detected between most system pairs that swapped ranks (using a paired-t-test with $\alpha = 5\%$). Even without correcting for non-significant system swaps, relatively high rank correlation of Kendall's τ and Spearman's ρ are obtained (cf. Table 4). We conclude that automatic relevance data is suitable for evaluating entity rankings.

Moreover, we find that the leaderboards of both the Wiki-18 and TQA subsets does not change much under the automatic test collection (cf. Figure 2e and 2f). The main difference is that system `C--1` moved by two ranks. In comparison, we observe many more system swaps under the corresponding manual leaderboards (cf. Figure 2b and 2c).

The automatic relevance data is based on entity links on input pages; which are also manually assessed. For TQA, 80% of positive automatic data was manually confirmed as relevant; 70% for Wiki-18. The discrepancy arises as the entity linking method does not distinguish between central and circumstantial entities. We suspect that when a large number of queries is used, all entity ranking systems are equally penalized by such false positive relevance assessments.

A large fraction of positive automatic data is rated by manual assessors with the highest relevance grade of "must be mentioned" (40% for TQA and 27% for Wiki-18). This means that of all confirmed relevant entities, about half are very central to the query.

4.6 RQ4: Entity Linking versus Edited Hyperlinks

Entity Linking algorithms such as Tagme provide possibly noisy annotations. Furthermore, an entity linking tools will link any detectable entity, while an editor writing an article would embellish the most informa-

Table 6: Two example entity rankings for the query facet "Zika fever/ Epidemiology" from middle (`A--2`) and the end (`D--1`) of the leaderboard of the Wiki-18 subset. Entities marked with "X" are relevant as automatically derived from the corresponding Wikipedia page section via entity linking (EL) or the hyperlinks included by the page editor (HL). While most entities are relevant for Zika fever, only few are relevant in the context of Epidemiology.

Rank	`A--2`	EL	HL	`D--1`	EL	HL
1	Guillain-Barre syndrome	X	X	Yellow fever	X	X
2	Microcephaly		X	Radial glial cell		
3	Zika virus	X		Mosquito-borne disease		
4	Dengue fever	X		Aedes africanus		
5	Yellow fever	X	X	Aedes apicoargenteus		
6	Fever			Zika virus	X	
7	Zika fever			Dengue fever		
8	Arthralgia			Zika virus outbreak timeline		
9	Conjunctivitis	X		Neonatal infection		
10	Vertically transmitted infection			2013-2014 Zika virus outbreaks in Oceania		

tive entities with hyperlinks to their Wikipedia pages. But editorial policies may influence the manually created hyperlinks in unexpected ways: Wikipedia editors only include a hyperlink at the first entity mention.

In Figure 3, we compare the entity ranking test collection derived with an entity linking toolkit versus hyperlinks that were manually included by the Wikipedia editors. This analysis is only conducted on the Wiki-18 subset of the entity ranking task, because the TQA subset does not include hyperlinks. While the absolute value of the evaluation measure is lower for editorial hyperlinks, we find that relative performance results are comparable overall. The most significant difference is that methods `C--1` and `C--2` moved from the middle of the field to the end of the leaderboard.

We use the manual assessments as a gold standard leaderboard and analyze the correlation of both automatic leaderboards with Kendall's τ and Spearman's rank correlation ρ. Results are presented in Table 5. We find that both entity linking and editorial hyperlinks correlate reasonably well with the manual assessments, where entity linking demonstrates a slightly lower correlation for MAP, which is possibly due to limitations of the manual benchmark with respect to recall. The strength of correlation between automatic and manual approaches is very similar to the correlation between both automatic test collections.

Table 6 provides entity rankings of two entity ranking systems for example query facet "Zika fever/ Epidemiology".[10] We see that the higher position of `A--2` on the leaderboard is due to a larger number of relevant entities in the top ranks. This is the case both when using entity links (EL) and hyperlinks from the page editor (HL). While most entities in both rankings are relevant for the page title "Zika fever", many are not specific for its epidemiology.

[10] `https://en.wikipedia.org/wiki/Zika_fever#Epidemiology`

5 Conclusion

This work examines an approach for automatic test collection creation that does not require any human assessments, which provides affordable access to large-scale test collections for passages and entities with useful additions, such as query facets and the possibility to derive a benchmark for entity-support passage retrieval or sub-topic clustering.

We demonstrate the validity of this approach for entity and passage ranking with the help of human assessors, who agree on the leaderboard of systems, obtaining Spearman's rank correlations that are consistently above 0.85. Furthermore, human assessors agree on the relevance of automatic passage relevance data which contains only 1% of false positives. However, the automatic test collection is stricter than the manually created benchmark, containing only about a third of positive assessments. We discuss and experimentally evaluate that this difference does not affect its reliability of distinguishing systems by rank quality. In contrast, the automatic benchmark provides the opportunity to study the marginal relevance of the ranking, by the nature of its construction. Another advantage of our approach is that it is less influenced by a particular selection of systems from which the assessment pool is built—a problem pointed out by Jayasinghe et al. (2014). Anecdotally, many participants reported that automatic collections are very effective for method development and machine learning, since performance on train and test splits are often nearly identical.

The experimental evaluation includes automatic test collections constructed from Wikipedia pages and middle school textbooks from the TQA corpus. We believe that this approach can also be applied to create test collections for many related tasks, such as entity support-passage retrieval and entity-based answer-passage retrieval. Our approach relies on the availability of a corpus of input pages, where titles and headings correspond to information needs and the content represents the desired response. A debate for future work is whether manual assessment time is better spent assessing pools or creating a corpus of suitable input pages.

We believe that this open-source test collection allows the IR community to gain a better understanding on how relevance is manifested in natural language. This understanding leads to better ad hoc retrieval models to which research on user models and interaction data should be applied. Our test collection approach was motivated by system evaluation rather than training. Nevertheless, the release of a very large automatic test collection for passage ranking made it possible to train data-hungry neural networks for this task (MacAvaney et al., 2019; Nogueira and Cho, 2019).

Acknowledgements

We express our gratitude for many suggestions of several experts in the field, such as Ellen Voorhees and Fernando Diaz, who helped to make this track successful. We are grateful for Ben Gamari's invaluable support in developing the test collection creation and assessment interface software. Finally, we thank all TREC CAR participants and reviewers for their valuable feedback.

This material is based upon work supported by the National Science Foundation under Grant No. 1846017. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Allan J (2003) Hard track overview in trec 2003: High accuracy retrieval from documents. In: TREC, pp 24–37

Alonso O, Mizzaro S (2012) Using crowdsourcing for TREC relevance assessment. Information Processing & Management 48

Arnold S, Schneider R, Cudré-Mauroux P, Gers FA, Löser A (2019) Sector: A neural model for coherent topic segmentation and classification. Transactions of the Association for Computational Linguistics 7:169–184

Arvola P, Kekäläinen J, Junkkari M (2010) Expected reading effort in focused retrieval evaluation. Information Retrieval 13(5)

Asadi N, Metzler D, Elsayed T, Lin J (2011) Pseudo test collections for learning web search ranking functions. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, ACM, pp 1073–1082

Azzopardi L, De Rijke M, Balog K (2007) Building simulated queries for known-item topics: an analysis using six european languages. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 455–462

Balog K, Neumayer R (2013) A test collection for entity search in DBpedia. In: Proceedings of the SIGIR, pp 737–740

Bast H, Buchhold B, Haussmann E, et al. (2016) Semantic search on text and knowledge bases. Foundations and Trends® in Information Retrieval 10(2-3):119–271

Bast H, Buchhold B, Haussmann E (2018) A quality evaluation of combined search on a knowledge base and text. KI-Künstliche Intelligenz 32(1):19–26

Beitzel SM, Jensen EC, Chowdhury A, Grossman D (2003) Using titles and category names from editor-driven taxonomies for automatic evaluation. In: Proceedings of the twelfth international conference on Information and knowledge management, ACM, pp 17–23

Berendsen R, Tsagkias M, De Rijke M, Meij E (2012) Generating pseudo test collections for learning to rank scientific articles. In: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, pp 42–53

Berendsen R, Tsagkias M, Weerkamp W, De Rijke M (2013) Pseudo test collections for training and tuning microblog rankers. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 53–62

Blanco R, Zaragoza H (2010) Finding support sentences for entities. In: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp 339–346

Bodoff D (2008) Test theory for evaluating reliability of IR test collections. Information Processing & Management 44(3):1117–1145

Boston C, Fang H, Carberry S, Wu H, Liu X (2014) Wikimantic: Toward effective disambiguation and expansion of queries. Data & Knowledge Engineering 90:22–37

Bota H, Zhou K, Jose JM, Lalmas M (2014) Composite retrieval of heterogeneous web search. In: Proceedings of the 23rd international conference on World wide web, ACM, pp 119–130

Callan JP (1994) Passage-level evidence in document retrieval. In: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, Springer-Verlag New York, Inc., pp 302–310

Chatterjee S, Dietz L (2019) Why does this entity matter? support passage retrieval for entity retrieval. In: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, ACM, pp 221–224

Choi E, He H, Iyyer M, Yatskar M, Yih Wt, Choi Y, Liang P, Zettlemoyer L (2018) QuAC: Question answering in context. arXiv preprint arXiv:180807036

Cormack GV, Palmer CR, Clarke CL (1998) Efficient construction of large test collections. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp 282–289

Dalton J, Dietz L, Allan J (2014) Entity query feature expansion using knowledge base links. In: Proceedings of the SIGIR, pp 365–374

Dalton J, Xiong C, Callan J (2019) CAsT 2019: The conversational assistance track overview. In: Text REtrieval Conference (TREC)

Dalvi B, Minkov E, Talukdar PP, Cohen WW (2015) Automatic gloss finding for a knowledge base using ontological constraints. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, ACM, pp 369–378

Demartini G, Iofciu T, De Vries AP (2009) Overview of the INEX 2009 entity ranking track. In: International Workshop of the Initiative for the Evaluation of XML Retrieval, pp 254–264

Dietz L (2019) ENT Rank: Retrieving entities for topical information needs through entity-neighbor-text relations. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, pp 215–224

Dietz L, Verma M, Radlinski F, Craswell N (2017) Trec complex answer retrieval overview. In: Text Retrieval Conference

Dietz L, Gamari B, Dalton J, Craswell N (2018) Trec complex answer retrieval overview. In: Text Retrieval Conference

Ferragina P, Scaiella U (2010) Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In: Proceedings of the 19th ACM international conference on Information and knowledge management, pp 1625–1628

Foley J, O'Connor B, Allan J (2016) Improving entity ranking for keyword queries. In: Proceedings of the CIKM, pp 2061–2064

Frank JR, Kleiman-Weiner M, Roberts DA, Voorhees E, Soboroff I (2014) Evaluating stream filtering for entity profile updates in trec 2012, 2013, and 2014 (kba track overview, notebook paper). In: Text Retrieval Conference (TREC)

Gabrilovich E, Ringgaard M, Subramanya A (2013) Facc1: Freebase annotation of clueweb corpora. Version 1:2013

Jayasinghe GK, Webber W, Sanderson M, Culpepper JS (2014) Improving test collection pools with machine learning. In: Proceedings of the 2014 Australasian Document Computing Symposium, p 2

Kamps J, Lalmas M, Pehcevski J (2007) Evaluating relevant in context: Document retrieval with a twist. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp 749–750

Kamps J, Geva S, Trotman A, Woodley A, Koolen M (2008) Overview of the INEX 2008 ad hoc track. In: International workshop of the Initiative for the Evaluation of XML Retrieval, Springer, pp 1–28

Kaszkiel M, Zobel J (1997) Passage retrieval revisited. In: ACM SIGIR Forum, vol 31, pp 178–185

Kembhavi A, Seo M, Schwenk D, Choi J, Farhadi A, Hajishirzi H (2017) Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In: Conference on Computer Vision and Pattern Recognition (CVPR)

MacAvaney S, Yates A, Cohan A, Soldaini L, Hui K, Goharian N, Frieder O (2019) Overcoming low-utility facets for complex answer retrieval. Information Retrieval Journal 22(3-4):395–418

Mendes PN, Jakob M, García-Silva A, Bizer C (2011) DBpedia spotlight: shedding light on the web of documents. In: Proceedings of the 7th international conference on semantic systems, ACM, pp 1–8

Nogueira R, Cho K (2019) Passage re-ranking with bert. arXiv preprint arXiv:190104085

O'Connor J (1980) Answer-passage retrieval by text searching. Journal of the American Society for Information Science 31(4):227–239

Pennington J, Socher R, Manning C (2014) GloVe: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543

Raviv H, Kurland O, Carmel D (2016) Document retrieval using entity-based language models. In: Proceedings of the SIGIR, pp 65–74

Sakai T, Song R (2011) Evaluating diversified search results using per-intent graded relevance. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, ACM, pp 1043–1052

Sawant U, Garg S, Chakrabarti S, Ramakrishnan G (2019) Neural architecture for question answering using a knowledge graph and web corpus. Information Retrieval Journal 22(3-4):324–349

Schuhmacher M, Dietz L, Paolo Ponzetto S (2015) Ranking entities for web queries through text and knowledge. In: Proceedings of CIKM, pp 1461–1470

Shah C, Pomerantz J (2010) Evaluating and predicting answer quality in community qa. In: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp 411–418

Soboroff I, Huang S, Harman D (2018) Trec 2018 news track overview. In: The Twenty-Seventh Text REtrieval Conference (TREC 2018) Proceedings

Voorhees EM (2001) Evaluation by highly relevant documents. In: Proceedings of the SIGIR, pp 74–82

Wade C, Allan J (2005) Passage retrieval and evaluation. Tech. rep., Massachusetts Univ Amherst Center for Intelligent Information Retrieval

Wan X, Yang J (2008) Multi-document summarization using cluster-based link analysis. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 299–306

Xiong C, Callan J (2015) EsdRank: Connecting query and documents through external semi-structured data. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, ACM, pp 951–960

Xiong C, Callan JP, Liu TY (2017a) Word-entity duet representations for document ranking. In: SIGIR

Xiong C, Power R, Callan J (2017b) Explicit semantic ranking for academic search via knowledge graph embedding. In: Proceedings of the 26th international conference on world wide web, International World Wide Web Conferences Steering Committee, pp 1271–1279

Yang Y, Yih Wt, Meek C (2015) WikiQA: A challenge dataset for open-domain question answering. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp 2013–2018

Yilmaz E, Kanoulas E, Aslam JA (2008) A simple and efficient sampling method for estimating AP and NDCG. In: Proceedings of the SIGIR, pp 603–610

Zhang H, Cormack GV, Grossman MR, Smucker MD (2018) Evaluating sentence-level relevance feedback for high-recall information retrieval. arXiv preprint arXiv:180308988