EXAM: How to Evaluate Retrieve-and-Generate Systems for Users Who Do Not (Yet) Know What They Want

David P. Sander[1], Laura Dietz[2]

[1]Bottomline Technologies, 325 Corporate Dr, Portsmouth, NH 03801, United States of America
[2]University of New Hampshire, Durham NH 03824, United States of America

Abstract

Our long-term goal is to develop systems that are forthcoming with information. To be effective, such systems should be allowed to combine retrieval with language generation. To alleviate challenges such systems pose for today's IR evaluation paradigms, we propose EXAM, an evaluation paradigm that uses held-out exam questions and an automated question-answering system to evaluate how well generated responses can answer follow-up questions—without knowing the exam questions in advance.

Keywords

Information Retrieval, Natural Language Generation, Evaluation, Conscious Information Needs

1. Introduction

Often users want to learn about a topic they know very little about. Taylor and Belkin [1, 2] call this a "conscious information need" originating from an "anomalous state of knowledge" where the user knows too little about the topic to ask precise questions. As a result, web search and conversational search systems do not provide a satisfying user experience. Instead, users often turn to Wikipedia. However, depending on the topic, articles may be out-of-date, incomplete, or missing. If this is the case, today's users embark on a journey of exploratory search where they are required to manually compile relevant information from multiple search requests.

As a remedy, research on interactive information retrieval is developing novel search interfaces [3]. We consider a complementary avenue by aiming to provide the best possible response in a single interaction turn, by compiling an overview for a topic of the user's choice.

Within the TREC Complex Answer Retrieval track [4], we aspire to retrieve-and-generate overview articles as found on Wikipedia. The objective of the third year of the track (CAR Y3) is to respond to the query with an article that is composed of existing paragraphs.

We offer a new evaluation based on whether these articles answer obvious follow-up questions. Examples are available in the online appendix.[1]

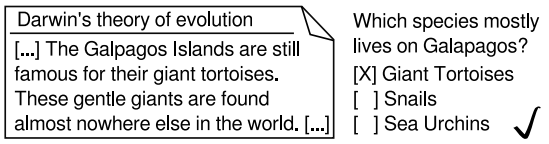

Figure 1: Evaluating articles (left) through EXAM questions.

1.1. Vision: Retrieve-and-Generate Systems

To compose a comprehensive overview article, our long-term goal is to develop retrieve-and-generate systems that automatically read the web and organize relevant information. An ideal overview article would educate the user about different aspects of the topic, with the goal of enabling the user to formulate precise questions or search queries. To not waste the user's time, the article should be forthcoming with relevant information that immediately answers obvious follow-up questions without being explicitly asked.

We envision such systems to perform retrieval, content planning, and natural language generation—all while inferring which pieces of information are relevant and how they fit together. We refer to such systems as *retrieve-and-generate systems* to indicate that retrieval is only the first step in the pipeline, and sources will be further processed—possibly using abstractive summarization or language generation—for presentation to the user.

GPT-3 [5], T5 [6], and other natural language generation (NLG) models offer a promising avenue for generating relevant text in combination with information retrieval (IR) systems. Recent models achieve great results with respect to grammar, flow of arguments, and readability. However, there are concerns whether these generated articles contain the most relevant information

DESIRES 2021 – 2nd International Conference on Design of Experimental Search & Information REtrieval Systems, September 15–18, 2021, Padua, Italy
dpsander42@gmail.com (D. P. Sander); dietz@cs.unh.edu (L. Dietz)
0000-0001-8508-6357 (D. P. Sander); 0000-0003-1624-3907 (L. Dietz)

[1]Appendix: https://www.cs.unh.edu/~dietz/appendix/exam/

[7]. Integrating NLG with retrieval will help to instill the user's trust in the faithfulness of the provided information. The development of retrieve-and-generate systems is hindered by the lack of an accepted evaluation paradigm for a fair comparison.

1.2. Evaluation Challenges

Typical IR evaluation paradigms are based on relevance assessments for texts from a known corpus in order to quantify the relevance of a ranking. The evaluation paradigm is directly applicable to predefined passages or "legal spans" [8, 9]. However, when arbitrary text spans can be retrieved or when retrieved text is modified, this evaluation paradigm needs to be adjusted. A common approach is to predict if retrieved text is sufficiently similar to assessed text, as in character-MAP [10, 11], passage-ROUGE [12], or BERTscore [13]. In a preliminary study we found BERTscore to be successful when the linguistic style is similar, e.g. both are Wikipedia paragraphs. However, when gold articles are linguistically different, BERTscore was found to be less reliable.

An open question is how to develop evaluation methods that can identify relevant *information* without being affected by the linguistic style in which the information is presented.

1.3. An Alternative Evaluation Approach

In this work, we discuss an alternative evaluation paradigm that directly assesses the usefulness of the retrieved *information*–instead of documents [14]. In particular, the evaluation paradigm is directly in service of our design goals: to educate the user about a topic they are not familiar with, while preemptively being forthcoming with answers.

We achieve this with a mostly[2] automatic metric, called the EXam Answerability Metric (EXAM). EXAM determines the quality of generated text by conducting an exam that assesses the article's suitability for correctly answering a set of query-relevant follow-up questions as depicted in Figure 1. Like an exam in school, the retrieve-and-generate systems must identify relevant information without knowing the exam questions beforehand. An external Q/A system will attempt to answer the follow-up questions; the more questions that can be answered correctly with the article, the higher the system's quality.

We suggest using exam questions that are relatively obvious follow-up questions to the user's request. For example, when a user provides a query such as "Darwin's Theory of Evolution", the generated comprehensive article should directly answer some reasonable follow-up questions such as "What species of bird did Darwin observe on the Galapagos Islands?" and, "Which scientists influenced Darwin's work?" The suggested evaluation paradigm assesses the system's ability to generate query-relevant articles which offer comprehensive information. The goal is to preempt the user with answers to potential follow-up questions, thus alleviating the user from the burden of asking obvious questions that could have been anticipated.

EXAM does not rely on relevance assessments or a fixed corpus. Once a sufficient question bank is created, it can be reused to evaluate future systems without any manual involvements. EXAM can compare retrieval-only systems as well as retrieve-and-generate systems. Since EXAM only assesses the information content, not the information source or document, it is a corpus independent metric that even allows the comparison of systems that use the open-web as a corpus or neural NLG systems.

1.4. The Chicken-and-Egg Problem

The development of novel retrieve-and-generate systems and the development of suitable evaluation paradigms form a chicken-and-egg problem: New systems (the "egg") cannot be studied without an established evaluation, while novel evaluation paradigms (the "chicken") cannot be tested without established retrieve-and-generate systems. With this work we provide the "chicken" by studying the efficacy of the EXAM evaluation metric on retrieval-only systems with respect to an established IR benchmark. The efficacy study uses systems submitted to the TREC Complex Answer Retrieval track in Year 3 for which a question bank is available through the TQA collection which was created by the Allen Institute for AI (AI2). We demonstrate that the leaderboard ranking of systems under EXAM correlates highly with the official track evaluation measures based on manual assessments created by the National Institute of Science and Technology (NIST). In contrast, using a collection of gold articles, we show that the system ranking under ROUGE does not correlate with the manual assessments, despite the fact that corresponding gold articles contain the right information and obtain a high EXAM score.

Contributions of this paper are as follows.

- Start a discussion on how to evaluate IR systems that further process retrieved raw text.
- Suggest EXAM, an alternative evaluation paradigm to complement existing evaluation strategies.
- Provide a study on TREC CAR Y3, by reusing exam questions from the related TQA data set. We demonstrate a high correlation with traditional IR metrics, even in cases where the linguistic style is too different for ROUGE to work.
- While our motivation arises from conscious information needs, the EXAM paradigm is applicable to many

[2]Fully automatic once a benchmark is created.

areas of IR, including ad hoc document retrieval and conversational search.

Outline. Section 2 provides an overview of related evaluation approaches. Section 3 introduces our EXAM evaluation paradigm. Section 4 outlines the experimental evaluation and discusses our results, before concluding in Section 5.

2. Related Work

2.1. Text Summarization

ROUGE [15] is one of the most popular evaluation metrics for text summarization, because the only human involvement is to create a reference summary. ROUGE and related metrics like METEOR [16] use n-gram overlap to quantify the similarity of phrases and vocabulary between two texts—one being the reference. Though ROUGE is commonly used, it has some drawbacks, as suggested by Scialom et al. [17] and Deutsch et al. [18]. Differing word choice between summaries results in lower ROUGE scores. Because of this, comparing two articles, by two different authors, both about the same topic, could have very low ROUGE scores due to dissimilar word choice. We use ROUGE-1 as a baseline for our evaluation paradigm in Section 4.

Alternatively, some automated metrics use a trained similarity to detect text that is classified as relevant, such as Reval [19], BERTScore [13], and NUBIA [20].

2.2. Summarization Evaluation with Q/A

Eyal et al. [21] compares the quality of generated summaries to a reference summary, using a Q/A system and questions generated from entities in the reference summary. Deutsch et al. [18] focuses on automatic question generation from a reference summary. Huang et al. [22] develop a CLOZE-style training signal that automatically derives multiple-choice questions from reference summaries. However, since not all phrases in a reference summary are equally important, such approaches cannot guarantee that the evaluation will test for *relevant* information.

Doddington [23] suggests evaluating machine translation by conducting an exam with questions. In 2002, these were answered by human annotators (not a Q/A system), however inconsistencies between annotators led to issues in the evaluation, which was then discontinued. We hypothesize that automatic Q/A systems (albeit not perfect) offer a fair comparison across systems.

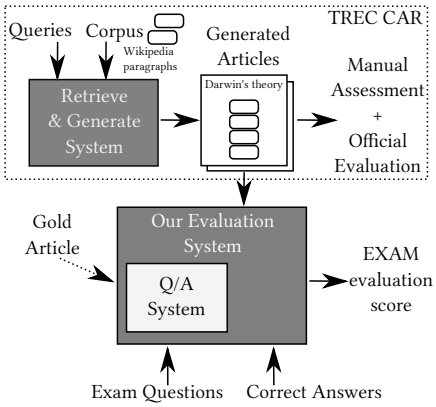

Figure 2: Pipeline of retrieve-and-generate systems and evaluation with our proposed evaluation paradigm.

2.3. Information Retrieval Evaluation

Information Retrieval is commonly evaluated with a pool-based Cranfield-style paradigm, where the top k documents are pooled and manually assessed for relevance [24]. Dietz and Dalton [8] automate manual IR assessment by deriving queries and a passage corpus from existing articles, then assess passages as relevant when they originated from the corresponding article and/or section. This method does not allow information to deviate from predefined passages.

To evaluate systems that retrieve passages of variable length, Keikha et al. [12] use a ROUGE metric to measure the n-gram overlap with ground truth passages. An alternative approach is to use a character-wise MAP to award credit for shared character sequences [10, 11].

In this work we discuss an alternative evaluation paradigm that focuses on retrieving *information*.

3. Approach: EXAM Evaluation

We propose the automatic EXam Answerability Metric (EXAM) for evaluating the usefulness of systems which retrieve and generate relevant articles in response to topical queries. These articles are evaluated based on how many exam questions an automated Q/A system can answer correctly. Our metric does not use relevance judgments nor reference summaries. Instead, a benchmark for EXAM consists of a set of queries with follow-up questions that a relevant article should answer. We use it to measure the relevance and completeness of comprehensive articles.

We first introduce our general evaluation approach, then explain customizations for evaluating CAR Y3 as used for evaluation.

3.1. EXAM Evaluation Paradigm

While motivated by conscious information needs, the evaluation paradigm can be applied to most topical IR tasks. Only a suitable bank of exam questions and a Q/A system need to be available.

3.1.1. Resources Required for Evaluation

We reserve exam questions and disallow the retrieve-and-generate systems under study to access them. Retrieve-and-read systems are given:

Queries: Given a free-text user query, such as "Darwin's Theory of Evolution", systems must generate a comprehensive response. Queries can be a simple keyword query or a more complex expression of information needs, such as a conversation prompt, desired sub-topics, or usage contexts.

The systems can access any corpus of their choice. EXAM does not require a predetermined corpus, unlike pool-based evaluations, such as the official TREC assessments [4]. Even corpus-free systems like GPT-3 can be evaluated.

Solely for the evaluation, we require the following resources to be available on a per-query basis:

Exam Questions and Answer Verification: A set of reasonably obvious follow-up questions about the query's topic. Any question style (e.g. multiple-choice, free text, etc.) can be used as long as the underlying Q/A system is trained to answer them and the answer can be automatically verified.

Q/A System: A high-quality Q/A system that is trained to answer exam questions. To be suitable, the Q/A system must use the given article to identify evidence for the question. All systems must be evaluated using the same Q/A system for EXAM scores to be comparable.

The evaluation process will use the above resources as depicted in Figure 2. We suggest using a multiple-choice Q/A system and an answer key to verify correctness. However, many Q/A systems can be used in our paradigm, such as the systems of Choi et al. [25], Nie et al. [26], or Perez et al. [27]. To be suitable, some Q/A systems would need to be customized, e.g. restricting the sentence selector of Min et al. [28].

3.1.2. EXAM Evaluation Scores

Given the queries and corpus, each system will generate one article per query. The exam questions are only used during evaluation: the Q/A system attempts to answer all exam questions based on the content of the generated article. For a query q, we measure the EXAM evaluation score of a generated article d_q from system S over the question bank for the query as,

$$\text{EXAM}(d_q|S) = \frac{\text{correct answers in } d_q}{\text{number of exam questions for } q}$$

The EXAM score of each retrieve-and-generate system is computed for each query (and hence generated article), then macro-averaged over multiple queries. Skipped queries are counted as zero score. EXAM awards no credit for unanswered or incorrectly answered questions, as these suggest that the generated article does not contain the right information.[3]

Similar to other proposals, e.g., of nugget-recall [29], EXAM is a recall-oriented evaluation measure. To penalize large amounts of non-relevant information, the article length can be restricted as in TREC CAR Y3, NTCIR One-click [30], or composite retrieval [31].

3.1.3. Normalizing EXAM with Gold Articles

As introduced above, our EXAM score enables relative quality comparison among systems and baselines. However, questions which are too difficult for the Q/A system to answer, or that are irrelevant to the query, could result in an artificially lowered score. To correct for this, we propose a relative-normalized EXAM score that uses human-edited gold articles d_q^* which are written to address the query and exam questions. This allows retrieve-and-generate systems to be scored using the context of an expected best-case scenario.

$$\text{n-EXAM}(S) = \frac{\sum_q \text{EXAM}(d_q|S)}{\sum_q \text{EXAM}(d_q^*)}$$

Note that if the gold articles contain less information than the generated articles, or are written obtusely, the retrieve-and-generated articles could earn a higher EXAM score than the gold articles, especially if the retrieve-and-generated articles express information clearer. This would result in an n-EXAM score above one. For example, a Q/A system would have difficulty extracting answers from a college textbook, when used as a gold article, as college-level reading material requires reader inference or significant logical deduction for full comprehension. This is one way that human-written, gold articles could receive lower EXAM scores than generated articles sourced from more plainly written corpora.

3.2. EXAM Evaluation for CAR Y3

The purpose of the CAR Y3 track is to study retrieval algorithms that respond to complex information needs

[3] Our focus is not on evaluating the Q/A system, but the usefulness of the generated article to a reader.

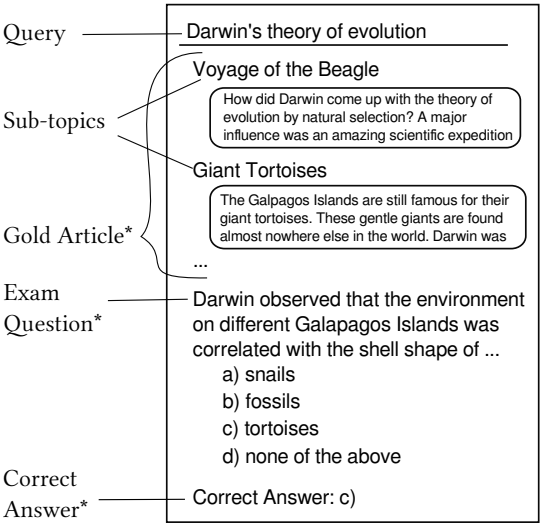

Figure 3: An example from the TQA dataset used to derive the TREC CAR benchmark as well as data for our EXAM evaluation measure (marked with *). The example is an excerpt from TQA entry L_0432/NDQ_009501.

by synthesizing longer answers by collating retrieved information, mimicking the style of Wikipedia articles. For the shared task to yield a reusable benchmark, participant systems were restricted to use a corpus of five million predefined passages without modifications.

The Textbook Question-Answering (TQA) [32] dataset provides textbook chapters and multiple-choice questions designed for middle school students. In CAR Y3, the queries were taken from titles of textbook chapters, and sub-topics were derived from headings. Sub-topics are used as "nuggets" in the official assessment and were provided to participants. For each query, participants were asked to produce relevant articles by selecting and arranging 20 paragraphs from the provided corpus of Wikipedia paragraphs.

For the example query "Darwin's Theory of Evolution" (tqa2:L_0432), an excerpt of a textbook chapter is depicted in Figure 3. Neither the questions nor the textbook content were available to the CAR Y3 participant systems. The figure depicts an example of how parts of the TQA dataset textbook chapters are used in the CAR Y3 benchmark versus held out for the EXAM evaluation. The connections between CAR Y3, retrieval systems under study, and our proposed evaluation paradigm are depicted in Figure 2.

3.2.1. Reference: Official CAR Y3 Evaluation

For selected queries (see Table 3), NIST assessors provided relevance annotations for all paragraphs in all submitted articles with respect to the query and sub-topics. Participants were encouraged to additionally submit paragraph rankings for each sub-topic. The official CAR Y3 evaluation was based on these rankings and the relevance assessments.[4]

3.2.2. Proposed Alternative: EXAM Evaluation

To evaluate the articles of participating systems with EXAM, we require a question bank of exam questions with an answer key. Queries are derived from titles of TQA textbook chapters which come with multiple-choice questions designed by the book author to test (human) students. We are using these multiple choice questions as a question bank for the EXAM metric to assess generated articles of participating systems. In particular, we use all provided non-diagram (i.e. not dependent on a picture) questions.

Gold Articles: We use the textbook content from the TQA textbook chapters as gold articles for the queries (also used by the ROUGE baseline as reference summary). While the EXAM metric does not require gold articles, we report the EXAM score achieved by the gold article for reference and include n-EXAM scores as well. As gold articles and questions were designed for middle school students, many answers are stated in an obtuse way and cannot be answered by simple text matches.

Used Question-Answering System: As a high-quality Q/A system we use the Decomposable Attention Q/A system provided by the organizers of the TQA challenge. The system is trained on the AI2 Reasoning Challenge dataset (ARC) [33]. The model is adapted from Parikh et al. [34], which performs the best on the SNLI [35] dataset, which contains questions similar to TQA questions.

As inputs, the Q/A system requires a text and a set of questions. The Decomposable Attention model searches the text for passages relevant to each question, then extracts answers by constructing an assertion per question and answer choice. Assertions without text support are eliminated, the most likely assertion under the attention model is returned as the answer. If all assertions are rejected, the question is not answered. Both unanswered and incorrectly answered questions result in a reduced EXAM score.

[4] Only manual assessments are available for CAR Y3. The automatic evaluation paradigm was only applicable to CAR Y1.

Table 1
Rank correlation between the leaderboards of different evaluation measures. Standard errors are below 0.02. Range: -1 to +1, higher is better.

	EXAM	Prec@R	MAP	nDCG20
ROUGE	-0.09	-0.01	-0.07	-0.01
nDCG20	0.74	0.94	0.95	
MAP	0.75	0.94		
Prec@R	0.74			

(a) Spearman's rank correlation coefficient.

	EXAM	Prec@R	MAP	nDCG20
ROUGE	-0.07	0.00	-0.05	0.00
nDCG20	0.57	0.86	0.88	
MAP	0.57	0.86		
Prec@R	0.56			

(b) Kendalls's tau rank correlation coefficient.

Table 3
Dataset statistics.

132	Queries with generated articles
20	Paragraphs per article per system
131	Queries with exam questions
2320	Exam questions
55	Queries with official TREC CAR assessments
303	Subtopics with official TREC CAR assessments
2790	Positively assessed paragraphs

4. Experimental Evaluation

We empirically evaluate EXAM as described in Section 3.2 using articles generated by the CAR Y3 [4] participant systems.

4.1. Experiment Setup

Due to the chicken-and-egg problem, no established retrieve-and-generate benchmarks exist and no established retrieve-and-generate systems are available. We base the experimental evaluation on sixteen retrieval systems submitted to CAR Y3. We use 131 queries that have a total of 2320 questions in the TQA dataset. We use each query's textbook chapter in the TQA dataset as a gold article. Dataset statistics are summarized in Table 3. Since these systems are not part of our work, we refer to the participant's description of their systems in the TREC Proceedings[5] and CAR Y3 Overview [4].

4.1.1. Evaluating the Evaluation Measure

Our goal is to find an alternative evaluation metric that—while mostly automatic—offers the same high quality as a manual assessment conducted by NIST. Hence, our measure of success is to produce a system ranking (i.e., leaderboard) that is highly correlated with the official CAR Y3 leaderboard. Low or anti-correlation suggests that an evaluation measure would not agree with a user's sense of relevance. Correlation of leaderboard rankings is measured in:

Spearman's Rank: High when each system S (of n) has a similar rank position under both leaderboards A, B: $\rho = 1 - \frac{6 \sum_S (rank_A(S) - rank_B(S))^2}{n(n^2-1)}$

Kendall's Tau: High when the rank order of many system pairs S_1, S_2 is preserved (P^+) versus swapped (P^-): $\tau = \frac{P^+ - P^-}{P^+ + P^-}$

Under any evaluation metric some systems obtain a similar evaluation score within standard error. As this is unlikely to indicate significant difference, we define such system pairs as tied, and thus attribute any score difference to random chance. Therefore, we randomly break ties, which is necessary for Spearman's rank, to produce the leaderboard and compute the rank correlation, repeating the process ten times. Results are presented in Tables 2a and 2b.

4.1.2. Metrics for System Quality

We study the leaderboard of systems under the following evaluation measures.

EXAM (ours): Our proposed evaluation measure which uses a Q/A system to evaluate generated articles (see Section 3.2).

n-EXAM (ours): A relative-normalized version of EXAM that uses a set of gold articles to contextualize the EXAM score.

Official CAR Y3 Evaluation (reference): Systems in CAR Y3 are evaluated using Precision at R (Prec@R), Mean-Average Precision (MAP), and Normalized Discounted Cumulated Gain at rank 20 (nDCG20) as implemented in trec_eval.[6]

ROUGE-1 F1 (baseline): ROUGE evaluates via the similarity between a generated article and a gold article. ROUGE-1 F1 combines precision and recall of predicting words in the summary. Words are lowercased and lemmatized, with punctuation and stopwords removed. We include ROUGE as a baseline evaluation paradigm, because it is fully automated and widely used in NLG.

[5]Proceedings: https://trec.nist.gov/pubs/trec28/trec2019.html

[6]TREC evaluation tool available here: https://github.com/usnistgov/trec_eval

Table 4
Quality of 16 participating systems submitted to TREC CAR Y3 as measured by our proposed EXAM, official TREC CAR metrics, and ROUGE. Systems are ordered by EXAM score, ranks under other metrics given in column "#", the best evaluation scores are marked in bold. Standard errors are about 0.01 or less. Systems whose performance is comparable to the gold articles are marked with "⋆" .

	Ours				Official TREC CAR Evaluation				Baseline	
Systems	EXAM	n-EXAM	#	⋆	Prec@R	MAP	nDCG20	#	ROUGE	#
rerank2-bert	**0.17**	**1.03**	1	⋆	0.22	0.18	0.31	3	0.42	12
dangnt-nlp	0.17	1.02	2	⋆	**0.28**	**0.25**	**0.38**	**1**	0.41	13
bert-cknrm-50	0.16	0.99	3	⋆	0.14	0.11	0.22	12	0.47	2
irit-run2	0.16	0.94	4	⋆	0.19	0.16	0.27	4	0.45	5
rerank3-bert	0.16	0.94	5	⋆	0.23	0.20	0.34	2	0.44	8
ict-b-convk	0.16	0.94	6	⋆	0.19	0.15	0.27	8	0.39	15
irit-run1	0.16	0.93	7	⋆	0.19	0.16	0.27	4	0.44	7
bm25-populated	0.15	0.93	8		0.18	0.14	0.25	9	0.43	10
unh-tfidf-ptsim	0.15	0.92	9		0.17	0.13	0.23	10	0.43	11
irit-run3	0.15	0.92	10		0.19	0.16	0.27	4	0.44	6
unh-bm25-ecmpsg	0.15	0.88	11		0.17	0.13	0.23	10	0.43	9
ecnu-bm25-1	0.14	0.87	12		0.19	0.15	0.27	7	**0.49**	**1**
ict-b-drmmtks	0.13	0.80	13		0.01	0.01	0.01	16	0.24	16
uvabottomupch.	0.09	0.57	14		0.04	0.03	0.06	14	0.45	4
uvabm25rm3	0.09	0.54	15		0.04	0.03	0.06	13	0.45	3
uvabottomup2	0.09	0.53	16		0.03	0.02	0.04	15	0.40	14
Gold articles (⋆)	0.17	1.00	-	-	-	-	-	-	-	-

4.2. Results

EXAM: Table 4 displays the leaderboard of tested retrieve-and-generate systems ordered by EXAM score. The trend is clear: Systems that rank high on the official TREC CAR leaderboard also rank high on the EXAM leaderboard, and systems that rank low on the official leaderboard also rank low on the EXAM leaderboard. For example, the rerank2-bert and dangnt-nlp participant systems are ranked as the top two systems by both the official leaderboard and the EXAM leaderboard. Some systems achieved similar EXAM scores as they likely produce the same passages ordered differently, affecting official TREC CAR evaluations but not EXAM. Additionally, due to the setup described in Section 3.2.2, the best systems even slightly surpass the gold articles (see discussion in Section 4.3). Many systems are performing within standard-error of the gold articles (marked with ⋆)—these are all similar systems based on the BERT neural language model.

Tables 2a and 2b display the Spearman's and Kendall's rank correlations of the EXAM and the official TREC CAR evaluation. Notably, EXAM achieves correlation of 0.74 in terms of Spearman's rank correlation and 0.56 in terms of Kendall's Tau, both values can range from -1 to 1. These averages illustrate just how strong EXAM correlates with official assessments, despite the much different evaluation paradigm. Prec@R, MAP, and nDCG20 use the same relevance assessments and evaluation paradigm. Hence it is not surpring that the correlation within official measures is higher than between EXAM and official measures.

ROUGE: By contrast, the leaderboard according to ROUGE-1 F1 is uncorrelated to the official TREC CAR leaderboard. Ecnu-bm25-1, which has the best ROUGE-1 F1 score, is not in the top five of the official leaderboard. Tables 2a and 2b demonstrate that the rank correlation of ROUGE is near zero across all metrics, which is equivalent to a random ordering.

4.3. Discussion

We discuss advantages and limitations of the paradigm.

Resilience to Q/A system errors: Any real-world Q/A system will make mistakes, most likely causing correct answers contained within the generated article to be missed. Indeed, the Q/A system is unable to correctly answer many questions with the gold article, in part because the article and questions were designed to be a challenge for middle school students. However, despite these Q/A errors, the study demonstrates EXAM reveals significant quality differences between systems. If our experiment had not been successful, we would not observe any correlation between EXAM and the official CAR Y3 assessments.

Overcoming linguistic differences: We found when using the gold article with ROUGE-F1, the system ranking does not agree with manual assessments. The issue originates from a difference in linguistic style, as generated articles are constrained to use Wikipedia paragraphs, but the gold articles are sourced from TQA. Hence, it is unlikely that gold articles would use the same phrases as the generated articles—despite both covering the same, relevant topics.

In a previous (unpublished) study on CAR Y2 data we found that ROUGE-F1 [12] obtains a reasonable correlation (Kendall's tau of 0.67, Spearman's rank of 0.67) when using manually assessed relevant paragraphs instead of gold articles. We conclude that ROUGE is struggling to overcome the linguistic differences between generated and gold articles.

In contrast, when evaluated under the EXAM measure, the gold article obtains the same score as the best participant system. Given the positive results, we conclude that EXAM is able to overcome the linguistic differences. We believe this is an important finding as the same issue is likely to arise when a retrieve-and-generate system uses external sources or a fully generative model.

Interpretation of N-EXAM: The n-EXAM metric can exceed 1.0 when the gold article is written obtusely, but the generated article explains relevant facts in accessible language. In our study, gold articles are designed for (human) students to carefully read the text and think about the answer, which is challenging for the Q/A system. In contrast, the submitted retrieval systems were allowed to select content from Wikipedia passages, which are likely to clearly state answers to obvious follow-up questions. Hence, we suggest to consider EXAM scores on gold articles as guidance, rather than a gold standard. Similar dataset biases are known from work on Multi-Hop Question Answering [36]. We remark that this issue also affects the ROUGE evaluation.

Universality of quality: Our evaluation paradigm is very different from pool-based Cranfield-style evaluations practiced in IR today [24]. Initial concerns that these paradigms evaluate different measures of quality have been ameliorated as the experimental evaluation demonstrates a high agreement between EXAM and the official CAR Y3 assessments.

Reduced manual effort: Cranfield-style evaluations involve a non-trivial amount of human labor. By contrast, EXAM's human assessors only develop a bank of questions that evaluate the information content of articles. EXAM question banks can be reused in a fully automatic manner, as the Q/A system conducts the exam.

While exam questions cannot test all possible useful follow-up questions, we demonstrate that the available question bank is large enough to measure significant differences between systems. To identify how little effort would still yield good results, we spent one hour to manually create ten questions. While error bars are larger, the results still correlate with the official leaderboard. (Study omitted due to space constraints).

Benchmark reusability and comparability: Many IR benchmarks mandate the use of unmodified elements from a fixed corpus. By contrast, EXAM uses a corpus-independent evaluation, and thus can be applied across different corpora and sources, including open web or NLG algorithms. Systems using different sources can all be evaluated and compared with each other using the EXAM evaluation.

5. Conclusions and Future Directions

We discuss an evaluation paradigm for retrieve-and-generate systems, which are systems that modify retrieved raw data before presentation. This poses a challenge for today's IR evaluation paradigms. To facilitate empirical research on retrieve-and-generate systems, we discuss an alternative evaluation paradigm, the EXam Answerability Metric (EXAM), that tests whether the system provides relevant information rather than the right documents.

EXAM uses a Q/A system and query-specific question banks to evaluate whether the system response is capable of answering some obvious follow-up questions, even without being explicitly asked to do so. We verify that leaderboards under the EXAM evaluation and the manual TREC CAR evaluation, agree with a Spearman's Rank correlation of 0.74 and Kendall's Tau of 0.56.

EXAM has two benefits over the traditional IR evaluation paradigm: it avoids the need for manual relevance assessments, and it can compare systems that use different (or no) corpora for retrieval. While gold articles and assessments can be used within the EXAM paradigm, at a minimum EXAM only requires humans to curate queries and question banks—the rest of the evaluation is fully automatic. EXAM also has an advantage over the text summarization metric, ROUGE [15], as EXAM evaluates documents by the relevance of information provided, rather than exact wording. This conclusion is in line with findings of Deutsch et al. [18].

While not studied in this work, EXAM could also be used to construct a training signal, as long as the exam questions are not available as inputs to the retrieve-and-generate system.

Our long-term goal is to develop systems to support

users who do not (yet) know what exactly they are looking for. We envision a system that synthesizes a comprehensive topical overview by collating retrieved text with post-processing steps like natural language generation. Permitting different linguistic styles and encouraging comprehensiveness, renders traditional IR evaluation paradigms as very costly. These goals also pose challenges regarding benchmark reuse for a fair comparison across systems. The EXAM evaluation paradigm provides a new avenue for retrieve-and-generate research to evaluate systems by information content.

However, EXAM can also evaluate many other information retrieval tasks: EXAM allows the comparison of ad hoc retrieval from fixed corpora with open-web retrieval. EXAM offers an alternative way to assess redundancy for search result diversification. EXAM can evaluate the information content of each turn of a conversational search system as well as the provided information content over multiple turns. We believe that in general, evaluation paradigms that, like EXAM, penalize avoidable conversation turns will encourage information systems that are forthcoming with answers.

Acknowledgments

We thank Peter Clark, Ashish Sabharwal, Tushar Khot from the Allen Institute for AI for their help with the TQA dataset and the provision of the Q/A System, and the UNH TREMA group for their guidance in doing this research.

References

[1] R. S. Taylor, Question-negotiation and information seeking in libraries, College & Research Libraries 29 (1968) 178–194.

[2] N. J. Belkin, Anomalous states of knowledge as a basis for information retrieval, Canadian journal of information science 5 (1980) 133–143.

[3] T. Ruotsalo, J. Peltonen, M. J. Eugster, D. Głowacka, P. Floréen, P. Myllymäki, G. Jacucci, S. Kaski, Interactive intent modeling for exploratory search, ACM Transactions on Information Systems (TOIS) 36 (2018) 1–46.

[4] L. Dietz, J. Foley, Trec car y3: Complex answer retrieval overview, in: Proceedings of Text REtrieval Conference (TREC), 2019.

[5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, arXiv preprint arXiv:2005.14165 (2020).

[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67.

[7] D. Ippolito, D. Duckworth, C. Callison-Burch, D. Eck, Automatic detection of generated text is easiest when humans are fooled, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 1808–1822.

[8] L. Dietz, J. Dalton, Humans optional? Automatic large-scale test collections for entity, passage, and entity-passage retrieval, Datenbank-Spektrum (2020) 1–12.

[9] W. R. Hersh, A. M. Cohen, P. M. Roberts, H. K. Rekapalli, Trec 2006 genomics track overview., in: TREC, volume 7, 2006, pp. 500–274.

[10] J. Kamps, M. Lalmas, J. Pehcevski, 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, ACM, 2007, pp. 749–750.

[11] C. Wade, J. Allan, Passage retrieval and evaluation, Technical Report, Massachusetts University Amherst Center for Intelligent Information Retrieval, 2005.

[12] M. Keikha, J. H. Park, W. B. Croft, Evaluating answer passages using summarization measures, in: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, ACM, 2014, pp. 963–966.

[13] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).

[14] J. Allan, B. Croft, A. Moffat, M. Sanderson, Frontiers, challenges, and opportunities for information retrieval: Report from swirl 2012 the second strategic workshop on information retrieval in lorne, in: ACM SIGIR Forum, volume 46, ACM New York, NY, USA, 2012, pp. 2–32.

[15] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://www.aclweb.org/anthology/W04-1013.

[16] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.

[17] T. Scialom, S. Lamprier, B. Piwowarski, J. Staiano, Answers unite! Unsupervised metrics for reinforced summarization models, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3246–3256. URL: https://www.aclweb.org/anthology/D19-1320. doi:10.18653/v1/D19-1320.

[18] D. Deutsch, T. Bedrax-Weiss, D. Roth, Towards question-answering as an automatic metric for evaluating the content quality of a summary, arXiv preprint arXiv:2010.00490 (2020).

[19] R. Gupta, C. Orasan, J. van Genabith, Reval: A simple and effective machine translation evaluation metric based on recurrent neural networks, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1066–1072.

[20] H. Kane, M. Y. Kocyigit, A. Abdalla, P. Ajanoh, M. Coulibali, Nubia: Neural based interchangeability assessor for text generation, arXiv preprint arXiv:2004.14667 (2020).

[21] M. Eyal, T. Baumel, M. Elhadad, Question answering as an automatic evaluation metric for news article summarization, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 3938–3948. URL: https://www.aclweb.org/anthology/N19-1395. doi:10.18653/v1/N19-1395.

[22] L. Huang, L. Wu, L. Wang, Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020). URL: http://dx.doi.org/10.18653/v1/2020.acl-main.457. doi:10.18653/v1/2020.acl-main.457.

[23] G. Doddington, Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, in: Proceedings of the Second International Conference on Human Language Technology Research, HLT '02, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002, p. 138–145.

[24] E. M. Voorhees, The evolution of cranfield, in: Information retrieval evaluation in a changing world, Springer, 2019, pp. 45–69.

[25] E. Choi, D. Hewlett, J. Uszkoreit, I. Polosukhin, A. Lacoste, J. Berant, Coarse-to-fine question answering for long documents, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 209–220.

[26] P. Nie, Y. Zhang, X. Geng, A. Ramamurthy, L. Song, D. Jiang, Dc-bert: Decoupling question and document for efficient contextual encoding, Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (2020). URL: http://dx.doi.org/10.1145/3397271.3401271. doi:10.1145/3397271.3401271.

[27] E. Perez, P. Lewis, W.-t. Yih, K. Cho, D. Kiela, Unsupervised question decomposition for question answering, arXiv preprint arXiv:2002.09758 (2020).

[28] S. Min, V. Zhong, R. Socher, C. Xiong, Efficient and robust question answering from minimal context over documents, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 1725–1735.

[29] J. Lin, Is question answering better than information retrieval? Towards a task-based evaluation framework for question series, in: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, 2007, pp. 212–219.

[30] T. Sakai, M. P. Kato, Y.-I. Song, Overview of ntcir-9, in: Proceedings of the 9th NTCIR Workshop Meeting, 2011, 2011, pp. 1–7.

[31] H. Bota, K. Zhou, J. M. Jose, M. Lalmas, Composite retrieval of heterogeneous web search, in: Proceedings of the 23rd international conference on World wide web, 2014, pp. 119–130.

[32] A. Kembhavi, M. Seo, D. Schwenk, J. Choi, A. Farhadi, H. Hajishirzi, Are you smarter than a sixth grader? Textbook question answering for multimodal machine comprehension, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 5376–5384.

[33] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, O. Tafjord, Think you have solved question answering? Try ARC, the AI2 reasoning challenge, ArXiv abs/1803.05457 (2018).

[34] A. Parikh, O. Täckström, D. Das, J. Uszkoreit, A decomposable attention model for natural language inference, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 2249–2255. URL: https://www.aclweb.org/anthology/D16-1244. doi:10.18653/v1/D16-1244.

[35] S. R. Bowman, G. Angeli, C. Potts, C. D. Manning, A large annotated corpus for learning natural language inference, arXiv preprint arXiv:1508.05326 (2015).

[36] J. Chen, G. Durrett, Understanding dataset design choices for multi-hop reasoning, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4026–4032.