Evaluating Answer Passages using Summarization Measures

Mostafa Keikha, Jae Hyun Park, W. Bruce Croft
CIIR, University of Massachusetts Amherst, Amherst, MA
{keikham, jhpark, croft } @cs.umass.edu

ABSTRACT

Passage-based retrieval models have been studied for some time and have been shown to have some benefits for document ranking. Finding passages that are not only topically relevant, but are also answers to the users' questions would have a significant impact in applications such as mobile search. To develop models for answer passage retrieval, we need to have appropriate test collections and evaluation measures. Making annotations at the passage level is, however, expensive and can have poor coverage. In this paper, we describe the advantages of document summarization measures for evaluating answer passage retrieval and show that these measures have high correlation with existing measures and human judgments.

1. INTRODUCTION

Some information retrieval (IR) queries can be best answered with a web page, others can be answered with a single fact or named entity. These types of queries, known as navigational and factoid questions, have been well-studied in the literature and the techniques for generating answers for them form the basis of many search engine result pages in both web and mobile environments. The category of informational queries is, however, very broad and many queries could potentially best be answered with a text passage that is longer than a factoid, but considerably shorter than a full web page. Passage retrieval has also been studied previously [3, 8, 2], but the main aim of this research was to improve the document ranking for a query by using passage-level evidence in combination with other features. Instead, our hypothesis is that there are queries for which a passage-level answer can be superior to a document-level answer and, for those queries, result lists that include passages will be more effective than documents alone.

In this paper, we focus on the critical issue of how to evaluate passage retrieval systems. Traditional document-level evaluation measures are not directly applicable to passage retrieval, because each variation of a passage retrieval model retrieves different passages. In order to use those document-level measures one would need to manually assess all possible passages, which is not practical. As an alternative, character-based measures have been developed that treat each character as a document and evaluate them using existing precision and recall measures [1, 5]. These character-based measures also have limitations and require the annotation of exact characters in order to consider it relevant. Annotating all the relevant passages in the large collections used currently is also not practical. On top of their limitations, these measures have been rarely studied for answer retrieval. Evaluating answer passages, as opposed to relevant passages, is even more difficult. Finding proper answers is usually more challenging and more ambiguous than finding relevant passages which makes answer annotation even more time consuming.

We employ summarization evaluation metrics for evaluating answer passage retrieval methods. These measures capture similarity of candidate passages to a sample of known or "ideal" answer passages. They do not require exhaustive annotation of passages, which makes them reasonable candidates for passage retrieval evaluation. Summarization measures address some important aspects of evaluation such as the amount of noise in a passage, size of a passage and the coverage over the ideal answers that are all crucial for a proper evaluation measure. We show that these measures are reasonably correlated with existing measures and human assessments. Further, we describe a cross-collection evaluation scenario as a new application of these measures that is not possible with existing metrics.

2. RELATED WORK

Passage analysis has been studied in the information retrieval community from different perspectives. Incorporating passages into document retrieval systems is the most common use of passage-level information [3, 8, 2]. Most of the work on passage retrieval uses passages as an intermediate representation for retrieving other types of objects such as documents. Less attention has been paid to directly retrieving passages instead of documents as final answers to a query. This problem was partly addressed in the HARD track in TREC, INEX ad hoc track and TREC Genomics track [1, 5, 4]. As part of these tasks, new evaluation measures were proposed. The proposed metrics, e.g., R-precision, use characters from annotated relevant passages that are found in the top retrieved passages for evaluating systems [1, 4]. The proposed character-level measures are generally similar to traditional document-level measures but use characters as opposed to documents. The 1-click task from the NTCIR workshop is similar but is more focused

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609485 .

on factual queries and requires an extra human annotation phase to extract the most important pieces of text (nuggets) for evaluation [6].

While we do not limit our study to any specific type of queries, we are interested in the situation where we have one or more ideal answers for the query and would like to compare the retrieved passages to the ideal answer. Thus we do not require any nugget annotation and the whole annotated answer is considered as a unit for evaluation. Content-based similarity measures were part of the HARD track initial evaluation proposal [1]. However, they were not employed in the official track evaluation and were not studied afterwards.

In order to compare a retrieved passage with an ideal annotated answer, we study the feasibility of using summarization measures. Previous studies show that summarization metrics are highly correlated with human judgments and can capture the quality of summaries [7]. In this paper, we study the behavior of those measures for evaluating answer passage retrieval and compare them to human judgments and existing measures. Further, we study the sensitivity of those measures to noisy judgments. Finally, we discuss an interesting application of these measures where we evaluate passages retrieved from one collection using judgments developed for another collection. This scenario that we call cross-collection evaluation is not possible using existing measures and has the potential to facilitate the creation of evaluation benchmarks.

3. ROUGE EVALUATION METRICS

Our evaluation is based on summarization evaluation metrics that are implemented in the ROUGE package [7]. Originally, the metrics were used to compare an automatically generated summary or translation against a reference or a set of reference (human-generated) summaries or translations. Essentially, any summarization metric compares two pieces of text based on their overlapping concepts. Depending on the type of concept, one can define different measures. We explore the following measures that are the most successful ones for the summarization task:

- ROUGE-N: a measure based on the N-gram co-occurrence statistics. We use N with values 1 and 2 that gives us unigram and bigram instances.
- ROUGE-S : this measure considers the number of overlapping skip-bigrams in the evaluation. A skip-bigram is any pair of words in the same order with a limited distance between them. We use ROUGE-S4 that considers any bigram with a distance less than 4.
- ROUGE-SU: this measure considers both unigrams and skip-bigrams in the evaluation. We use ROUGE-SU4 that considers any bigram with a distance less than 4.

ROUGE-1, ROUGE-2 and ROUGE-SU4 are the official evaluation metrics used in the Document Understanding Conference (DUC) for evaluating summaries [7]. We perform stemming and stop word removal before comparing the two passages. In any of the ROUGE measures, the number of overlapping concepts can be compared to the total number of concepts in the retrieved passage or in the ideal answer or combination of the two that results in a precision, recall and F1 variation of each measure [7]:

$$\text{ROUGE-x}_{precision} = \frac{|IdealAnswer_x \bigcap RetrievedPassage_x|}{|RetrievedPassage_x|}$$

$$\text{ROUGE-x}_{recall} = \frac{|IdealAnswer_x \bigcap RetrievedPassage_x|}{|IdealAnswer_x|}$$

$$\text{ROUGE-x}_{F1} = \frac{2 \times \text{ROUGE-x}_{precision} \times \text{ROUGE-x}_{recall}}{\text{ROUGE-x}_{precision} + \text{ROUGE-x}_{recall}}$$

x shows the concept type and has values $1, 2, S4$ and $SU4$ representing four different concept types. $Passage_x$ shows the set of x concepts extracted from the $Passage$. As we can see, each of these measures calculates one value for each passage.

The recall variation captures what portion of the relevant concepts (concepts in the ideal answer) is present in the retrieved passage. On the other hand, the precision variation of each measure captures what portion of the retrieved concepts are among the relevant concepts. The F1 measure combines the precision and recall values and gives a single evaluation value for each passage [7]. Analogous to the size of the ranked list in the traditional IR metrics, with increasing size of the retrieved passage we generally expect recall to increase and precision to decrease.

When there are multiple ideal answers for a query, each retrieved passage will have multiple evaluation values; one value for each ideal answer. We explored two options for aggregating these values that includes averaging and maximum value. Our experiments showed that maximum value is a better choice for passage evaluation. Due to lack of space we only show the results of maximum function here.

After evaluating each single retrieved passage with respect to the query, we need to aggregate the the results in order to evaluate the ranked list as a whole. The simplest option would be to average the passage-level evaluation values over all the retrieved passages. We also explored other options in which we give more importance to the higher ranks, similar to the nDCG measure in document retrieval. While the nDCG variation improves the evaluation performance, the difference is not significant and for the sake of clarity and space we do not report those results in this paper.

Finally we average the evaluation values over all the queries to get the performance evaluation of a system.

4. COMPARISON TO EXISTING MEASURES

Character-level measures are the most comparable existing measures to the ROUGE measures. These measures have been studied before in the context of HARD track and Genomics track in TREC and the ad-hoc track in INEX [1, 5, 4]. They use a set of highlighted passages as relevant answers and treat each of the highlighted characters as a relevant item. Then they calculate existing evaluation measures such as precision and recall over characters.

In our first experiments, we assume that character-based measures are good evaluation measures and examine the correlation between them and ROUGE measures. We get a set of submitted runs to INEX and evaluate them using the official INEX evaluation toolkit. Then for all the submitted runs we extract their real text and evaluate them using ROUGE measures. Finally, we estimate the correlation between system performance using INEX measure and

Table 1: Correlation between ROUGE measures and character-based MAP

Measure	Precision	Recall	F1
ROUGE-1	-0.01	0.69	0.61
ROUGE-2	0.57	0.70	0.69
ROUGE-S4	0.51	0.33	0.61
ROUGE-SU4	0.53	0.65	0.59

ROUGE measures. After cleaning and removing systems with non-valid passages, we have 39 systems to evaluate.

Table 1 shows the correlation between ROUGE measures and the character-based MAP measure. As we can see, except for ROUGE-1 that is based on the unigram overlap, the rest of the measures have quite high correlation with the character-based measure. The best measure is ROUGE-2 that is based on bigram overlap and we can see that all its variations have high correlation.

These results show that ROUGE measures, while having other benefits that we will discuss later, can provide a similar ranking of systems to existing character-based measures.

5. COMPARISON TO HUMAN JUDGMENT

In the previous section, we assumed that character-based MAP is an ideal measures and having high correlation with it is a desired requirement. Given that this assumption is not necessarily true, in the next experiments we compare the ROUGE measures directly to human judgments.

For this analysis, we built a data set using the GOV2 collection and the corresponding TREC queries. Three human annotators including one graduate student and two undergraduate students were involved in the annotation process. The undergraduate students performed the main annotation task and the graduate student controlled the annotation results to make sure they are in a proper format and contain meaningful passages. We divided topics randomly in two different groups, one for each annotator. For each topic, we retrieved the top 50 documents using the Sequential Dependence Model (SDM), a state-of-the-art retrieval model. From the retrieved documents, we selected the relevant documents, based on the TREC relevance judgments, for the passage annotation phase. Each assessor annotated all the documents related to the topics assigned to him. Further, in order to study agreement between annotators, they also annotated the top five documents for the rest of the topics.

Annotators were asked to use our annotation toolkit and highlight all the answer passages in the document set. An answer passage is defined as a piece of text in a document that can answer the user information need. Our annotation guideline considers different properties of passages including how complete is the answer with respect to the query and how much non-relevant information it contains. Based on these criteria, we defined four level of answers as "perfect", "excellent", "good" and "fair". A perfect answer means the passage provides all the necessary information to answer the query and does not contain any non-relevant information. A fair answer provides some information regarding the query but it does not completely answer the query or it contains noise. Good and excellent answers are better that fair and worse than perfect answers. It is worth noting that all these answers are reasonable and the difference between them is generally marginal.

Our assessor found answer passages for 82 TREC queries and highlighted 8,027 passages, which is about 97 passages per query on average. Among all the annotated passages 43% of them are perfect answers, 44% are excellent, 10% are good and the rest are fair answers.

Table 2: Correlation between ROUGE measures and human judgment

Measure	Precision	Recall	F1
ROUGE-1	0.38	0.47	0.43
ROUGE-2	0.43	0.47	0.45
ROUGE-S4	0.42	0.46	0.44
ROUGE-SU4	0.42	0.47	0.45

Our annotators highlighted 84,381 words in the passage answers. Among these words, 59,693 of them are highlighted by one annotator, 46,660 of them by other annotator and 21,972 of them are highlighted by both annotators. Considering non-highlighted words in the judged documents as negative examples, the term-level kappa ratio between our annotators is about 0.38. This is comparable to answer-level agreement in previous studies where kappa value is reported about 0.3 [10].

In the next analysis, our goal is to test if ROUGE measures can distinguish between passage answers with different relevance levels. Since we have a graded relevance level for each passage, we can compare those grades to their scores assigned by ROUGE measures. For each query, we select some of the "perfect" passages with probability 0.5 as our ideal set of answers and we end up with about 20 ideal answers per query on average. We then evaluate the rest of the annotated passages using ROUGE measures by comparing them to the ideal answers. To this end, we assign a numeric value of 1,2,3,4 to "fair", "good", "excellent" and "perfect" grades respectively. We then calculate the correlation between these values and ROUGE outputs.

Table 2 shows the correlation values. As we can see, there is no significant difference between measures in this experiment. All the measures are comparably correlated with human judgments and all the correlation are at a statistically significant level with p-values less than 0.05. This shows that ROUGE measures can reasonably indicate the quality of the passages. It is worth noting that, the difference between different levels of relevance in our annotated passages, e.g. perfect and excellent, are very minimal. We believe that distinguishing between non-relevant and relevant passages would be even easier and ROUGE measures would perform better in a general evaluation scenario.

When we compare the content of a retrieved passage to a set of ideal answers, the quality of ideal answers is very important factor. In the previous experiment, we sampled the ideal answers only from the "perfect"-labeled answers. In the next experiments, we study the sensitivity of ROUGE measures to noisy judgments where we have also non-perfect answers, e.g. excellent answers, as part of the ideal set.

To this end, we randomly select a subset of the excellent answers with probability 0.5 and add them to the ideal set. Again we evaluate the rest of the passages by comparing them to the ideal set using ROUGE measures. Table 3 shows the correlations values when we have noisy judgments. We can see that all the correlation values are decreased, which shows the ROUGE measures are in fact sensitive to the quality of the ideal answers. However, the correlations are reasonably high and all of them are at statistical significant levels with p-values less than 0.05.

6. CROSS-COLLECTION EVALUATION

Table 3: Correlation between ROUGE measures and human judgment with noisy judgments

Measure	Precision	Recall	F1
ROUGE-1	0.30	0.41	0.35
ROUGE-2	0.35	0.40	0.37
ROUGE-S4	0.34	0.40	0.36
ROUGE-SU4	0.34	0.40	0.36

Table 4: Correlation between ROUGE measure and manual evaluation on CQA data

Measure	Precision	Recall	F1
ROUGE-1	0.41	0.30	0.28
ROUGE-2	0.56	0.32	0.31
ROUGE-SU4	0.51	0.32	0.30

As opposed to existing character-based measures, ROUGE measures do not compare exact positions in the documents for evaluation. This property enable us to use annotation from one document collection to evaluate passages that are retrieved from another collection.

In order to explore the feasibility of this option, we study the question answering problem. In a Community Question Answering (CQA) environment, questions and answers are provided by the users. Some users ask a question and other users answer the question or vote to the already provided answers. One of the interesting tasks in such environment is to automatically answer new questions. Evaluating the output of such system is a challenging task that might need a lot of annotations. In this section, we study if ROUGE measures can eliminate the need for annotation by directly comparing retrieved passages with the best human answers.

We use the Yahoo CQA data and manually select a set of fifty questions whose best answer is a coherent piece of text and has a chance to be found in our web collection. We generate a query based on each question by stemming and removing stop words. We use the resulted queries to retrieve passages from the Clueweb-B collection using the built-in passage retrieval functionality in Indri. We retrieve fixed size passages with length 50 terms and overlap 25 as shown to be effective choice [2].

We then use the best human-provided answer for each question as our ideal answer and evaluate the retrieved passages using ROUGE measures. Further we selected 200 top retrieved passages and manually assigned a relevance score to them between 0 and 4, where 0 is a non-relevant answer and 4 means a perfect answer. We then calculate the correlation between the manual assessments and the ROUGE values. The results are shown in table 4. In all the cases the correlation is at a statistically significant level (p-value less than 0.05). Again we can see that ROUGE measures can reasonably capture the quality of answer passages. As a more detailed inspection, Figure 1 shows the distribution of ROUGE-SU4 precision values for different levels of human judgments. We can clearly see those passages with high value of human assessment (level 3 or 4) have higher values for the ROUGE measures as well.

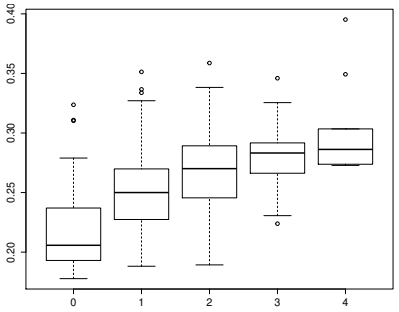

Figure 1: Distribution of REUGE-SU4 precision values for different levels of manual assessment.

7. CONCLUSION AND FUTURE WORK

In this paper, we investigated the evaluation of the answer passage retrieval task where the goal is to retrieve small passages, as opposed to full documents, in response to a user query. We employed text summarization evaluation metrics and showed that they are reasonably correlated with existing measures and human judgments. This suggests that ROUGE measures are reasonable measures to use, in addition to the existing measures, for evaluating passages.

Based on this new evaluation framework, our future work will be more focused on the passage-specific retrieval models. Due to the short length of passages, incorporating NLP features and translation models in the retrieval system seems to be a promising direction.

8. ACKNOWLEDGMENT

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF IIS-1160894. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

9. REFERENCES

[1] J. Allan. Hard track overview in trec 2004 - high accuracy retrieval from documents. In *proceedings of TREC*, 2004.

[2] M. Bendersky and O. Kurland. Utilizing passage-based language models for document retrieval. In *proceedings of ECIR*, pages 162–174, 2008.

[3] J. Callan. Passage-level evidence in document retrieval. In *proceedings of SIGIR*, pages 302–310, 1994.

[4] W. R. Hersh, A. M. Cohen, P. M. Roberts, and H. K. Rekapalli. Trec 2006 genomics track overview. In *Proceedings of TREC'06*, 2006.

[5] J. Kamps, J. Pehcevski, G. Kazai, M. Lalmas, and S. Robertson. Inex 2007 evaluation measures. In *proceedings of INEX workshop*, pages 24–33, 2007.

[6] M. P. Kato, T. Sakai, T. Yamamoto, and M. Iwata. Report from the ntcir-10 1click-2 japanese subtask: baselines, upperbounds and evaluation robustness. In *proceedings of SIGIR*, pages 753–756, 2013.

[7] C.-Y. Lin and F. J. Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *proceedings of ACL*, pages 605–612, 2004.

[8] X. Liu and W. Croft. Passage retrieval based on language models. In *proceedings of CIKM*, pages 375–382, 2002.

[9] T. Sakai, D. Ishikawa, N. Kando, Y. Seki, K. Kuriyama, and C.-Y. Lin. Using graded-relevance metrics for evaluating community qa answer selection. In *proceedings of WSDM*, pages 187–196, 2011.