Towards a Better Metric for Evaluating Question Generation Systems

Preksha Nema[1,2] **Mitesh M. Khapra**[1,2]
[1]Department of Computer Science and Engineering, Indian Institute of Technology Madras
[2] Robert Bosch Center for Data Science and Artificial Intelligence,
Indian Institute of Technology Madras
{preksha,miteshk}@cse.iitm.ac.in

Abstract

There has always been criticism for using n-gram based similarity metrics, such as BLEU, NIST, *etc*, for evaluating the performance of NLG systems. However, these metrics continue to remain popular and are recently being used for evaluating the performance of systems which automatically generate questions from documents, knowledge graphs, images, *etc*. Given the rising interest in such automatic question generation (AQG) systems, it is important to objectively examine whether these metrics are suitable for this task. In particular, it is important to verify whether such metrics used for evaluating AQG systems focus on *answerability* of the generated question by preferring questions which contain all relevant information such as question type (Wh-types), entities, relations, *etc*. In this work, we show that current automatic evaluation metrics based on n-gram similarity do not always correlate well with human judgments about *answerability* of a question. To alleviate this problem and as a first step towards better evaluation metrics for AQG, we introduce a scoring function to capture *answerability* and show that when this scoring function is integrated with existing metrics, they correlate significantly better with human judgments. The scripts and data developed as a part of this work are made publicly available.[1]

1 Introduction

The advent of large scale datasets for document Question Answering (QA) (Rajpurkar et al., 2016; Nguyen et al., 2016; Joshi et al., 2017; Saha et al., 2018a) knowledge base driven QA (Bordes et al., 2015; Saha et al., 2018b) and Visual QA (Antol et al., 2015; Johnson et al., 2017) has enabled the development of end-to-end supervised models for QA. However, as is always the case, data-hungry neural network based solutions could benefit from even more training data, especially in specific domains which existing datasets do not cater to. Creating newer datasets for specific domains or augmenting existing datasets with more data is a tedious, time-consuming and expensive process. To alleviate this problem and create even more training data, there is growing interest in developing techniques that can automatically generate questions from a given source, say a document (Du et al., 2017; Du and Cardie, 2017), knowledge base (Reddy et al., 2017; Serban et al., 2016), or image (Li et al., 2017) . We refer to this task as Automatic Question Generation (AQG). For example, given the document in Table 1, the task is to automatically generate a question whose answer is also contained in the document.

Document: In **1648** before the term "genocide" had been coined , the Peace of Westphalia was established to protect ethnic, racial and in some instances religious groups.
Possible Question: In which year was the Peace of Westphalia established ?

Table 1: A sample question generated by a human.

Given the practical importance of AQG and its potential to influence research in QA, it is not surprising that there has been prolific work in this field in the past one year itself (Jain et al., 2017; Li et al., 2017; Zhang et al., 2017; Du et al., 2017; Duan et al., 2017). Before this field grows further, it is important that the community critically examines the current evaluation metrics being used for this task. In particular, there is a need to closely examine the utility of existing n-gram based similarity metrics such as BLEU (Papineni et al., 2002), METEOR (Lavie and Denkowski,

[1]https://github.com/PrekshaNema25/Answerability-Metric

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3950–3959
Brussels, Belgium, October 31 - November 4, 2018.

2009), NIST (Doddington, 2002), *etc.* which have been adopted for this task. This work is a first step in that direction where we propose that apart from n-gram similarity, any metric for AQG should also take into account the *answerability* of the generated questions. With the help of a few examples below, we illustrate that *answerability* depends on the presence of relevant information such as question type (Wh-types), entities, relations, *etc*, and it is possible that a generated question has a high BLEU score but is still unanswerable and hence not useful.

To begin with, consider the task of answering questions from a Knowledge Base. Let us assume that the intended (gold standard) question is "Who was the director of Titanic?" and two different AQG systems generate the following questions "S1: director of Titanic?" and "S2: Who was the director of?". Any n-gram based evaluation metric would obviously assign a higher score to S2 (BLEU3: 81.9) than S1 (BLEU3: 36.8). However, as should be obvious S1 contains all the relevant information, and most humans would be easily able to understand and answer this question. A good evaluation metric should capture this notion of *answerability* and give more importance to relevant words in the question which brings us to the question "Which words are relevant?"

The above example might give the impression that *named entities* are essential but other words are not. However, this is misleading and may not always be the case. For example, consider these questions over an image: "Are the cats drinking milk?" v/s "How many cats are drinking milk?". These two questions have very different meaning indicating that even words like *are* and *how* are also crucial. Similarly, consider the task of answering questions from a passage titled "Matt Damon". In this case, most humans will be able to answer the question "What is the birth date of" even though the named entity is missing given that the passage only talks about "Matt Damon". Thus, in some cases, depending on the source (document, knowledge base, image) different portions of the question may be important.

To concretize the intuitions developed with the help of the above examples, we first collect human judgments. Specifically, we take questions from existing datasets for document QA, knowledge base QA and visual QA and add systematic noise to these questions. We show these questions to humans and ask them to assign scores to these questions based on the *answerability* and hence the usefulness of these questions (*i.e.*, whether the question contains enough information for them to be able to answer it correctly). We also compute various n-gram similarity metrics (BLEU, METEOR, NIST) comparing the noisy questions to the original questions and show that these metrics do not correlate well with human judgments. Similar studies (Callison-Burch et al., 2006; Liu et al., 2016) have already shown that these metrics do not correlate well with *fluency, adequacy, coherence* but in this work, we focus on *answerability*.

Based on the human evaluations, we propose to modify existing metrics to focus on *answerability* in addition to n-gram similarity. The idea is to make these metrics flexible such that, if needed, the weight assigned to *answerability* and n-gram similarity can be adjusted depending on the task (document QA, Knowledge-Base QA, Visual QA). Further, for capturing *answerability* we propose additional weights for different components of the question (question type, content words, function words, and named entities) These weights can be learned from a small amount of human annotated data and may differ from task to task.

2 Related Work

We have organized our literature survey into 2 parts: (i) question generation systems (ii) studies which analyze evaluation metrics used for NLG.

Question Generation: Early work on question generation used rule-based approaches to generate questions from declarative sentences (Heilman and Smith, 2010; Mostow and Chen, 2009; Lindberg et al., 2013; Labutov et al., 2015). More recent works use attention based neural models for question generation (Du and Cardie, 2017; Du et al., 2017). Some models (Yuan et al., 2017) feed the generated questions to a QA system and use the performance of the QA system as an indicator of the quality of the questions. A few models (Wang et al., 2017; Tang et al., 2017) treat question answering (QA) and question generation (QG) as complementary tasks and focus on jointly training for these two tasks. Other models focus only on the performance of the QA task (Yang et al., 2017; Duan et al., 2017) and not explicitly on the quality of the generated questions. Apart from generating questions from text there is also research on gen-

erating questions from images (Jain et al., 2017; Li et al., 2017; Zhang et al., 2017) and knowledge base (Serban et al., 2016; Reddy et al., 2017).

Evaluation metrics for NLG: Current popular metrics for NLG such as BLEU (Papineni et al., 2002), METEOR(Lavie and Denkowski, 2009), ROUGE (Lin, 2004) and NIST (Doddington, 2002) essentially compute the n-gram similarity between the reference sentence and the generated sentence. Though these metrics are very popular and are used for a wide range of NLG tasks including AQG, there has always been criticism for using these metrics (for example, see (Callison-Burch et al., 2006; R et al., 2007; Callison-Burch, 2009)). More recently, there has been criticism (Liu et al., 2016) for using such metrics for evaluating dialog systems eventually resulting in a new metric (Lowe et al., 2017). This new metric while very important, came a bit late in the day and much after several dialog systems were proposed, evaluated and compared using the above n-gram based metrics. It is very important to prevent a similar situation in question generation where many systems get proposed without evaluating them using the right metric. Our work is a first step in this direction, and we hope it will lead to more research in designing the right metrics for AQG.

3 Current Evaluation Metrics

We give a quick overview of the metrics which are currently used for evaluating AQG systems.

BLEU: BLEU is a precision-based evaluation metric which considers exact n-gram matches. For a given value of n, the precision is computed as the fraction of n-grams in the generated hypothesis which match some n-gram in the reference hypothesis. The final BLEU score is computed as the geometric mean of the n-gram precisions obtained by varying n from 1 to N where N is typically 3 or 4. It also contains a brevity penalty to penalize hypothesis that are too short.

METEOR: As opposed to BLEU, METEOR uses both precision and recall, *i.e.*, it computes the fraction of the hypothesis which matches the reference (precision) as well as the fraction of the reference which is contained in the hypothesis (recall). Further, unlike BLEU which only considers exact matches, METEOR also considers matches with stemmed words, synonyms, and paraphrases. It also gives different weightage to matches corresponding to function words and matches corresponding to content words. The final score is the harmonic mean of the precision and recall calculated based on these four matches. Additionally, it also includes a fragmentation penalty to account for gaps and differences in word order. In effect, METEOR is a parametric metric where the different parameters, *viz.*, (i) fragmentation penalty, (ii) weights of different matchers (exact, stemmed, synonyms, paraphrases) and (iii) weights of function and content words, are tuned to maximize correlation with human judgments.

NIST: NIST is a variant of the standard BLEU metric that takes into account the relative importance of each n-grams in the sentence. In particular, the metric gives a high weightage to n-grams which have a lower frequency in the corpus and hence are more informative as compared to very frequent n-grams which are less informative. Further, unlike BLEU which takes the geometric mean of n-gram precisions, NIST takes the arithmetic mean of these precisions. Additionally, they make a small change to the brevity penalty to minimize the impact of minor variations in the length of the hypothesis.

ROUGE: ROUGE is a set of evaluation metrics which were proposed in the context of automatic summarization. Typically, most studies use ROUGE-L, which is F-measure based on the Longest Common Subsequence (LCS) between a candidate and target sentence. Given two sequences, a common subsequence is the set of words which appear in both the sequences in the same order but unlike n-grams the common subsequence does not need to be contiguous. LCS is the longest of such common subsequences. For example, given the sentences candidate:"the boy went home" and reference:"the boy will go home", "the boy home" is the longest common subsequence even though it is not contiguous.

4 Human Judgments For Answerability

As mentioned earlier, for AQG, in addition to n-gram similarity, we also need to focus on the *answerability* of the generated questions. As illustrated in Section 1, answerability of a question depends on whether it contains all relevant information, such as question type (Wh-types), named entities and content words (often relations). Further, depending on the task (document QA, knowledge-base QA or visual QA) the importance of these words may vary. We perform human evaluations

to ascertain the importance of each of these components across different QA tasks. These evaluations allow us to independently analyze the importance of each of these components for the 3 QA tasks. In the remainder of this section, we describe the (i) process of creating noisy questions (ii) instructions given to the evaluators and the (iii) inferences drawn from human evaluations.

4.1 Creating Noisy Questions

We took 1000 questions each from 3 popular QA datasets, *viz.*, SQuAD, WikiMovies, and VQA. SQuAD (Rajpurkar et al., 2016) is a reading comprehension dataset consisting of around 100K questions based on passages from around 500 Wikipedia articles. The WikiMovies dataset contains around 100K questions which can be answered from a movie knowledge graph containing 43K entities and 9 relations *(director, writer, actor, etc.)*. The VQA dataset is an image QA dataset containing 265,016 images with around 5.4 questions on average per image.

We then created noisy versions of these questions using one of the following four methods:

Dropping function words: We refer to the list of English function words as defined in NLTK (Loper and Bird, 2002) and drop all such words from the question. Note that a noisy question with all function words dropped will have a very low BLEU score compared to the original question.

Dropping Named Entities: In our setup, identifying named entities in questions was easy because the questions were well formed and all named entities were capitalized. Alternately, we could have used the Stanford NER. However, on manual inspection, we found that marking the capitalized words as named entities were sufficient. We randomly dropped at most three named entities per question. This allows us to study how humans rate the output of an AQG system which does not contain the correct named entities.

Dropping Content Words: Words other than function words and named entities are also crucial for *answerability*. For example, "Who killed Jane?" and "Who married Jane?" lead to totally different answers. The word "killed/married" is very relevant to ascertain the correct answer. These words typically capture the relation between the entities involved (for example, *killed (John, Jane)*). We identify such important (content) words as ones which are neither question types (7-Wh questions) nor named entities nor stop-words. This perturbation allows us to study how humans rate an AQG system which does not produce the correct content (relation) words.

Changing the Question type: Changing the question type can lead to a different answer altogether or can make the question incoherent. For example the answers to "Who killed Jane?" and "What killed Jane?" are completely different. We create a noisy question by randomly changing the type of the question (for example. replace "who" with "what"). These question types are well defined (7-Wh questions including "*how*") and hence it is easy to identify and replace them. This allows us to study the importance of correct question type in the output of an AQG system.

Note that, an alternate way of collecting human judgments would have been to take the output of existing AQG systems and ask humans to assign answerability scores to these questions based on the presence/absence of the above mentioned relevant information. However, when we asked human evaluators to analyze 200 questions generated by an existing AQG system, they reported that the quality was poor. In particular, after having discussions with annotators, we found that using this output, it would be very difficult to conduct such a systematic study to assess the importance of different words in the question. Hence, we chose to use systematically simulated noisy questions.

4.2 Instructions

We asked the annotators to rate the *answerability* of the above noisy questions on a scale of 1-5. The annotators were clearly told whether the questions belonged to documents or knowledge bases or images. In our initial evaluations, we also tried showing the actual source (image or document) to the annotators. However, we realized that this did not allow us to do an unbiased evaluation of the quality of the questions. The annotators inferred missing information from the document or image and marked the question as answerable (even though the relevant entity *cat* is missing in the question). For example, consider the image of a cat drinking milk and the question "What is the drinking ?" If a human is shown the image then she can easily infer that the missing information is "cat" and hence mark the question as answerable. This clearly biases the study, and therefore we did not show the source to the evaluators.

Rating	Description	Examples
1	All important information is missing and it is impossible to answer the question	"What is against the ~~sign~~ ?", "Why is using ~~O2~~ instead of ~~CO2~~ less efficient?"
2	Most of the important information is missing and I can't infer the answer to the question	"Which films did ~~Lee H. Katzin~~ direct ?", "Low doses of ~~anti-inflammatories~~ are sometimes used with what ~~classes~~ of drugs?"
3	Some important information is missing leading to multiple answers	"What Harvard Alumni ~~was the~~ Palestine Prime Minister?", "What country ~~is the~~ teaching subject discussing?"
4	Most of the important information is present and I can infer the answer	"How ~~far~~ from the Yard is the Quad located?","what ~~films~~ did Melvin Van Peebles star in?"
5	All important information is present and I can answer the question	"What globally popular half marathon began ~~in~~ 1981?", "What kind ~~of~~ vehicle ~~is~~ parked ~~the~~ sidewalk?"

Table 2: Instructions along with the examples. The striked out words were removed as a part of systematic noise from the original question.

Dataset	κ	Pearson	Spearman
SQuAD	0.63	0.823	0.795
WikiMovies	0.81	0.934	0.927
VQA	0.70	0.842	0.822

Table 3: Inter annotator agreement, Pearson and Spearman coefficients between Human Scores.

Metric	SQuAD		WikiMovies		VQA	
	Pearson	*Spearman*	*Pearson*	*Spearman*	*Pearson*	*Spearman*
BLEU1	0.167	0.165	0.179	0.144	-0.025*	-0.048*
BLEU2	0.100*	0.103*	0.072*	0.087*	-0.075*	-0.091*
BLEU3	0.080*	0.086*	0.036*	0.001*	-0.126	-0.114
BLEU4	0.065*	0.067*	-0.020*	-0.011*	-0.086*	-0.127
ROUGE-L	0.165	0.158	0.091*	0.043*	-0.009*	-0.053*
METEOR	0.107	0.124	0.198	0.214	-0.035*	0.009*
NIST	0.173	0.158	0.088*	-0.033*	0.158	0.169

Table 4: Correlation between existing metrics and human judgments. Note that the values with $*$ are **not** statistically significant (p-value > 0.01).

A total of 25 in-house annotators participated in our study, and we got each question evaluated by *two* annotators. The annotators were Computer Science graduates competent in English. We did an initial pilot using the instructions mentioned in Table 2, but due to the subjective nature of the task, it was difficult for the annotators to agree on the notion of *important information*. In particular, we found that the annotators disagreed between *most important information* and *all important information* (*i.e.*, they were confused between rating 1 v/s 2 and 4 v/s 5). We, therefore, did a small pilot with a group of 10 annotators and asked them to evaluate around 30 questions from each dataset and help us refine the guidelines to define the notion of importance clearly. Based on group discussions with the annotators we arrived at additional example based guidelines to help them distinguish between cases where *"all the"*, *"most of the"* and *"some of the"* important information is present. The original instructions and various examples (some of which are shown in Table 2) were then shared and explained to all the annotators, and they used these to provide their judgments.

4.3 Human-Human Correlation

In Table 3, we report the average inter-annotator agreement between the ratings using Cohen's kappa (κ) score (Cohen, 1968). Based on guidelines in (McHugh, 2012) we note that we have a strong inter-annotator agreement for WikiMovies and moderate agreement for SQuAD and VQA. Figure 1 indicates that there is a linear correlation between the two ratings for each question and hence we measured the correlation using Pearson coefficient. For completeness, we also measure the monotonic correlation using Spearman coefficient. The Spearman coefficient is slightly lower than the Pearson coefficient because the inter-annotator agreement is stronger at the tail of the distribution *i.e.*, when the question is either very bad (Rating: 1) or very good (Rating: 5).

4.4 Correlation between human scores and existing evaluation metrics

We first compute BLEU, METEOR, NIST and ROUGE-L score for each noisy question by comparing it to the original question. We then compute the correlation of each of these scores with annotator ratings. Note that to compute correlation, the annotator ratings are combined to obtain a gold score. The ratings are normalized using the normalization method mentioned in (Blatz et al., 2004) and then averaged to obtain the gold score. For SQuAD and VQA, we observe that NIST which gives more weightage to informative n-grams correlates better than other metrics. For WikiMovies, METEOR which even allows non-

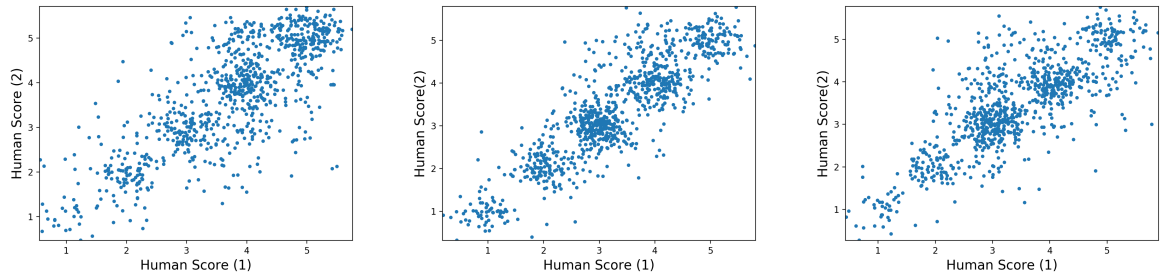

Figure 1: Human-Human Correlation for SQUAD, WikiMovies and VQA respectively.

exact word matches correlates better than other metrics. For SQuAD and WikiMovies, the correlation of human scores with the simple unigram based BLEU1 score is higher than that with other metrics. This is in line with the observation we made earlier that humans can understand and answer questions that are not well-formed, *e.g.*, "What birth-date Damon?".

5 Modifying existing metrics for AQG

The above study suggests that existing metrics do not correlate well with human judgments about *answerability*. We propose modifications to these metrics so that in addition to n-gram similarity they also account for *answerability*. Based on the human evaluations, we found that *answerability* mainly depends on the presence of 4 types of elements, *viz.*, *relevant content words*, *named entities* and *question types* and *function words*. As outlined in Section 4.1 it is easy to identify these elements in the question. Let $c(S_r), c(S_n), c(S_q)$ and $c(S_f)$ be the number of **r**elevant words, **n**amed entities, **q**uestion words and **f**unction words respectively in the noisy question which have corresponding matching words in the gold standard reference question. We can then compute the weighted average of the precision and recall of each of these elements as

$$P_{avg} = \sum_i w_i \frac{c(S_i)}{|l_i|} \qquad R_{avg} = \sum_i w_i \frac{c(S_i)}{|r_i|}$$

where $i \in \{r, n, q, f\}$, $\sum_i w_i = 1$ and $|l_i|$, $|r_i|$ is the number of the words belonging to i^{th} type of element in the noisy question and reference sentences respectively. Just to be clear r, n, q, f stand for *relevant content words*, *named entities* and *question types* and *function words* respectively. Note that w_i's are tunable weights and in Section 5.1, we explain how to tune these weights.

Datasets	w_{ner}	w_{imp}	w_{sw}	w_{qt}	δ
SQuAD	0.41	0.36	0.03	0.20	0.66
WikiMovies	0.55	0.31	0.02	0.11	0.83
VQA	0.04	0.59	0.15	0.21	0.75

Table 5: Coefficients learnt for Q-BLEU1 from human judgments across different datasets.

$$\text{Answerability} = 2.\frac{P_{avg}R_{avg}}{P_{avg} + R_{avg}}$$

We can combine this *answerability* score with any existing metric (say, BLEU4) to derive a modified metric for AQG as shown below:

$$Q\text{-BLEU4} = \delta\text{Answerability} + (1-\delta)\text{BLEU4} \quad (1)$$

such that $\delta \in \{0, 1\}$ to make sure that Q-Metric ranges between 0 to 1. Similarly, we can derive Q-NIST, Q-METEOR and so on.

5.1 Tuning the weights w_i's and δ

We tuned the weights (w_i's and δ) using the human annotation data. For each source (document, knowledge-base, and images), annotators evaluated 1000 noisy questions. The annotator scores were first scaled between 0 to 1 using the normalization method in (Blatz et al., 2004), and the normalized scores were averaged to obtain the final gold score. For each source, we used 300 of these annotations and used bagging to find the optimal weights. In particular, we drew 200 samples randomly from the given set of 300 samples and did a grid search to find w_i's and δ such that the Q-METRIC computed using Equation 1 had maximum correlation with human scores. We repeated this process for $k = 20$ times and computed the optimal w_i's and δ each time. We found that for

Q-Metric	SQuAD		WikiMovies		VQA	
	Pearson	*Spearman*	*Pearson*	*Spearman*	*Pearson*	*Spearman*
Q-BLEU1	0.258	0.255	0.828	0.841	0.405	0.384
Q-BLEU2	0.244	0.243	0.825	0.835	0.390	0.360
Q-BLEU3	0.239	0.240	0.824	0.837	0.374	0.331
Q-BLEU4	0.233	0.232	0.826	0.837	0.373	0.311
Q-ROUGE-L	0.253	0.249	0.821	0.841	0.402	0.385
Q-METEOR	0.158	0.157	0.821	0.837	0.402	0.378
Q-NIST	0.246	0.248	0.824	0.845	0.384	0.346

Table 6: Correlation between proposed Q-Metric and human judgments. All the correlations have a p-value < 0.01 and hence statistically significant.

any given weight (w_i) the standard deviation was very low across these k experiments. For each w_i and δ we obtained the final value by taking an average of the values learned in each of the k experiments. We also observed that the weights did not change much even when we used more data for tuning. Also note that we tuned these weights separately for each metric (*i.e.*, Q-BLEU4, Q-NIST, Q-METEOR and so on). For illustration, we report these weights for Q-BLEU1 metric in Table 5. As expected, the weights depend on the source from which the question was generated. Note that for WikiMovies, named entities have the highest weight. For VQA content words are most important, as they provide information about the entity being referred to in the question. Note that for SQuAD and VQA, the original base metric also gets weightage comparable to other components, indicating that a fluent question makes it easier to understand thus making it answerable. The overall trend for the values of w_i's was similar for other Q-METRICs also (*i.e.*, for Q-NIST, Q-METEOR and so on).

5.2 Correlation between Human scores and different Q-METRICs

Once the weights are tuned, we fix these weights and compute the Q-METRIC for the remaining 600-700 examples and report the correlation with human judgments for the same set of examples (see Table 6). For a fair comparison, the correlation scores reported in Table 4 are also on the same 600-700 examples. The correlation scores obtained for different Q-METRICs are indeed encouraging. In particular, we observe that while the correlation of existing metrics with noisy questions generated was very low (Table 4), the correlation of the modified metrics is much higher. This suggests that adding the learnable component for *answerability* and tuning its weights indeed leads to a better-correlated metric. Note that for VQA and SQuAD the correlations are not as high as human-human correlations, but the correlations are still statistically significant. We acknowledge that there is clearly scope for further improvement and the proposed metric is perhaps only a first step towards designing an appropriate metric for AQG. Hopefully, the human evaluation data released as a part of this work will help to design even better metrics for AQG.

5.3 Qualitative Analysis

We have listed some examples in Table 7, which highlight some strengths and weakness of the proposed Q-METRIC. We categorize examples as positive/negative depending on the similarity between human scores for answerability and the Q-BLEU score. For the examples marked as positive, the Q-BLEU score is very close to the *answerability* score given by humans.

6 Extrinsic evaluation

So far we have shown that existing metrics do not always correlate well with human judgments and it is possible to design metrics which correlate better with human judgments by including a learnable component to focus on *answerability*. We would now like to propose an extrinsic way of evaluating the usefulness of the proposed metric. The motivation for this extrinsic evaluation comes from the fact that one of the intended purposes of the modified metrics is to use them for training QA systems. Suppose we use a particular metric for evaluating the quality of an AQG system and suppose this metric suggests that the questions generated by this system are poor. We would obviously

Dataset		Original Question	Modified Question	Human Scores	QBLEU
SQuAD	**Positive**	What is another type of accountant other than a CPA? In addition to schools, where else is popularly based authority effective?	What is another type of accountant other than a ? In addition schools, where else popularly based authority effective?	0.10 0.85	0.47 0.83
	Negative	When did Tesla begin working for the Continental Edison Company? What famous person congratulated him?	When did begin working for the Continental Edison Company? What person congratulated him?	0.10 0.85	0.84 0.17
VQA	**Positive**	What color is the monster truck? What is in the polythene ?	What color monster truck? What is in the ?	0.92 0.10	0.81 0.14
	Negative	Why there are no leaves on the tree? How are the carrots prepared in the plate?	Why are leaves the tree? How carrots prepared plate?	0.35 0.10	0.73 0.68
WikiMovies	**Positive**	what films does Ralf Harolde appear in ? what is a film directed by Eddie Murphy ?	what films Ralf Harolde appear ? Which a film directed by Eddie Murphy ?	0.97 0.91	0.91 0.88
	Negative	what films does Gerard Butler appear in ? John Conor Brooke appears in which movies ?	how does Gerard Butler appear in ? appears in which movies ?	0.15 0.03	0.89 0.44

Table 7: Human (Gold) and Q-Metric scores for some of the examples from the collected human-evaluation data.

Type of Noise	BLEU	QBLEU	Hit 1
None	100	100	76.5
Stop Words	25.4	84.0	75.6
Question Type	74.0	79.3	73.5
Content Words	29.4	64.3	54.7
Named Entity	41.9	48.5	17.97

Table 8: Performance obtained by training on different types of noisy questions (WikiMovies).

Noise	BLEU	QBLEU	F1
None	100	100	76.5
Question Type	80.1	66.1	69.0
Stop Words	24.2	61.0	70.4
Content Words	60.7	57.1	64.1
Named Entity	77.0	56.0	73.8

Table 9: Performance obtained by training on different types of noisy questions (SQuAD).

Noise	BLEU	QBLEU	Acc(%)
None	100	100	64.4
Content Words	49.4	58.2	60.21
Question Type	63.7	50.9	59.81
Stop Words	10.8	37.7	57.37

Table 10: Performance obtained by training on different types of noisy questions (VQA).

discard this system and not use the questions generated by it to train a QA system. However, if the metric itself is questionable, then it is possible that the questions were good enough, but the metric was not good to evaluate their quality. To study this effect, we create a noisy version of the training data of SQuAD, WikiMovies, and VQA using the same methods outlined in Section 4.1. We then train a state of the art model for each of these tasks on this noisy data and evaluate the trained model on the original test set of each of these datasets. The models that we considered were (Seo et al., 2016) for SQuAD, (Miller et al., 2016) for WikiMovies and (Ben-younes et al., 2017) for VQA.

The results of our experiments are summarized in Table 8 - 10. The first column for each table shows the manner in which the noisy training data was created. The second column shows the BLEU4 score of the noisy questions when compared to the original reference questions (thus it tells us the perceived quality of these questions under the BLEU4 metric). We consider BLEU4 because of all the current metrics used for AQG it is the most popular. Similarly, the third column tells us the perceived quality of these questions under the Q-BLEU4 metric. Ideally, we would want that the performance of the model should correlate better with the perceived quality of the training questions as identified by a given metric. We observe that the general trend is better *w.r.t.* the Q-BLEU4 metric than the BLEU4 metric (*i.e.*, in general, higher Q-BLEU4 indicates better performance and lower Q-BLEU4 indicates poor performance). In particular, notice that BLEU4 gives much importance to stop words, but these words hardly have any influence on the final performance. We believe that such an extrinsic evaluation should also be used while designing better metrics and it would help us get better insights.

7 Conclusion

The main aim of this work was to objectively examine the utility of existing metrics for AQG. Specifically, we wanted to see if existing metrics account for the *answerability* of the generated questions. To do so, we took noisy generated questions from three different tasks, *viz.*, document QA, knowledge base QA and visual QA, and showed that the *answerability* scores assigned by humans did not correlate well with existing metrics. Based on these studies, we proposed a modification for existing metrics and showed that with the proposed modification these metrics correlate better with human judgments. The proposed modification involves learnable weights which can be tuned (depending on the source) using the human

judgments released as a part of this work. Finally, we propose an extrinsic evaluation with the aim of assessing the end utility of these metrics in selecting good AQG systems for creating training data for QA systems. Though the proposed metric correlates better with human judgments, there is still scope for improvement especially for document QA and visual QA. As future work, we would like to design better metrics for answerability and check if a non-linear combination of different elements in the Q-Metric leads to better correlation with human judgments.

8 Acknowledgements

We would like to thank Google for supporting Preksha Nema through their Google India Ph.D. Fellowship Program. We would also like to express our gratitude to the volunteers who participated in human evaluations.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In International Conference on Computer Vision (ICCV).

Hedi Ben-younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. Mutan: Multimodal tucker fusion for visual question answering. In The IEEE International Conference on Computer Vision (ICCV). volume 1, page 3.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, page 315.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. CoRR abs/1506.02075.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1. Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '09, pages 286–295.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of bleu in machine translation research. In EACL. The Association for Computer Linguistics.

J Cohen. 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological bulletin 70(4):213220.

George Doddington. 2002. 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, pages 138–145.

Xinya Du and Claire Cardie. 2017. Identifying where to focus in reading comprehension for neural question generation. In EMNLP. Association for Computational Linguistics, pages 2067–2073.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In ACL (1). Association for Computational Linguistics, pages 1342–1352.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In EMNLP. Association for Computational Linguistics, pages 866–874.

Michael Heilman and Noah A. Smith. 2010. Good question! statistical ranking for question generation. In HLT-NAACL. The Association for Computational Linguistics, pages 609–617.

Unnat Jain, Ziyu Zhang, and Alexander G. Schwing. 2017. Creativity: Generating diverse questions using variational autoencoders. In CVPR. IEEE Computer Society, pages 5415–5424.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In CVPR. IEEE Computer Society, pages 1988–1997.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In ACL (1). Association for Computational Linguistics, pages 1601–1611.

Igor Labutov, Sumit Basu, and Lucy Vanderwende. 2015. Deep questions without deep understanding. In ACL (1). The Association for Computer Linguistics, pages 889–898.

Alon Lavie and Michael J. Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. Machine Translation 23(2-3):105–115.

Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, and Xiaogang Wang. 2017. Visual question generation as dual task of visual question answering. CoRR abs/1709.07192.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Proc. ACL workshop on Text Summarization Branches Out. page 10.

David Lindberg, Fred Popowich, John C. Nesbit, and Philip H. Winne. 2013. Generating natural language questions to support learning on-line. In ENLG. The Association for Computer Linguistics, pages 105–114.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In EMNLP. The Association for Computational Linguistics, pages 2122–2132.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1. ETMTNLP '02, pages 63–70.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In ACL (1). Association for Computational Linguistics, pages 1116–1126.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. Biochemia medica: Biochemia medica 22(3):276–282.

Alexander H. Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In EMNLP. The Association for Computational Linguistics, pages 1400–1409.

Jack Mostow and Wei Chen. 2009. Generating instruction automatically for the reading strategy of self-questioning. In AIED. IOS Press, volume 200 of Frontiers in Artificial Intelligence and Applications, pages 465–472.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016..

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In ACL. ACL, pages 311–318.

Ananthakrishnan R, Pushpak Bhattacharyya, M Sasikumar, and Ritesh M Shah. 2007. Some issues in automatic evaluation of english-hindi mt: More blues for bleu. In International Conference on Natural Language Processing.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016. pages 2383–2392.

Sathish Reddy, Dinesh Raghu, Mitesh M. Khapra, and Sachindra Josh. 2017. Generating natural language question-answer pairs from a knowledge graph using a RNN based question generation model. In EACL (1). Association for Computational Linguistics, pages 376–385.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018a. Duorc: Towards complex language understanding with paraphrased reading comprehension. In ACL (1). Association for Computational Linguistics, pages 1683–1693.

Amrita Saha, Vardaan Pahuja, Mitesh M. Khapra, Karthik Sankaranarayanan, and Sarath Chandar. 2018b. Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph. In AAAI. AAAI Press.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. CoRR abs/1611.01603.

Iulian Vlad Serban, Alberto García-Durán, Çaglar Gülçehre, Sungjin Ahn, Sarath Chandar, Aaron C. Courville, and Yoshua Bengio. 2016. Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. In ACL (1). The Association for Computer Linguistics.

Duyu Tang, Nan Duan, Tao Qin, and Ming Zhou. 2017. Question answering and question generation as dual tasks. CoRR abs/1706.02027.

Tong Wang, Xingdi Yuan, and Adam Trischler. 2017. A joint model for question answering and question generation. CoRR abs/1706.01450.

Zhilin Yang, Junjie Hu, Ruslan Salakhutdinov, and William W. Cohen. 2017. Semi-supervised QA with generative domain-adaptive nets. CoRR abs/1702.02206.

Xingdi Yuan, Tong Wang, Çaglar Gülçehre, Alessandro Sordoni, Philip Bachman, Saizheng Zhang, Sandeep Subramanian, and Adam Trischler. 2017. Machine comprehension by text-to-text neural question generation. In Rep4NLP@ACL. Association for Computational Linguistics, pages 15–25.

Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, Jianfeng Lu, and Anton van den Hengel. 2017. Asking the difficult questions: Goal-oriented visual question generation via intermediate rewards. CoRR abs/1711.07614.