QSTS: A Question-Sensitive Text Similarity Measure for Question Generation

Sujatha Das Gollapalli and See-Kiong Ng
Institute of Data Science
National University of Singapore
{idssdg,seekiong}@nus.edu.sg

Abstract

While question generation (QG) has received significant focus in conversation modeling and text generation research, the problems of comparing questions and evaluation of QG models have remained inadequately addressed. Indeed, QG models continue to be evaluated using traditional measures such as BLEU, METEOR, and ROUGE scores which were designed for other text generation problems. We propose QSTS, a novel Question-Sensitive Text Similarity measure for questions that characterizes their target intent based on question class, named-entity, and semantic similarity information.

We show that QSTS addresses several shortcomings of existing measures that depend on n-gram overlap scores and obtains superior results compared to traditional measures on publicly-available QG datasets. We also collect a novel dataset SimQG for enabling question similarity research in QG contexts. SimQG contains questions generated by state-of-the-art QG models along with human judgements on their relevance with respect to passage contexts as well as the given reference questions. Using SimQG, we showcase the key aspect of QSTS that differentiates it from all existing measures. QSTS is not only able to characterize similarity between two questions, but is also able to score questions with respect to passage contexts. Thus QSTS is, to our knowledge, the first metric that enables the measurement of QG performance in a reference-free manner.

1 Introduction

Automatic Question Generation (QG), the task of generating natural language questions for a given input text passage continues to garner significant research focus in the NLP community (Wang et al., 2020b; Huang et al., 2021) due its potential application in education (Srivastava and Goodman, 2021), tutoring (Lindberg et al., 2013) and interactive dialog systems (Wang et al., 2020a).

In current research, in lieu of human evaluation, the standard practice for evaluating the performance of QG models involves the use of Question Answering (QA) datasets containing pairs of (reference question, passage context) elements. For evaluating QG, the machine-generated question for a given passage context is compared with the given reference question by applying metrics such as BLEU (Papineni et al., 2002), METEOR Lavie and Agarwal (2007), and ROUGE (Lin, 2004).

The above widely-used measures were originally developed for evaluating tasks such as summarization and translation and are based on overlap of n-grams between a given reference text and the model-generated text. Though studies have indicated that these measures do not correlate well with human judgements of fluency, relevance, and coherence (Callison-Burch et al., 2006; Liu et al., 2016; Nema and Khapra, 2018) these measures are easy to compute and continue to be used for various natural language generation (NLG) tasks. Recently though, research studies are addressing metrics learning for NLG using transformers and these learnt metrics were shown to obtain state-of-the-art performance in evaluation (Zhang et al., 2020; Sellam et al., 2020).

We posit that the existing metrics for measuring text generation tasks are inadequate for comparing questions due to their inability to incorporate various features that characterize questions. The first among these features is the question class (alternatively referred to as answer type) which places constraints on the answer to a given question.1 For example, for the question, “Who was Lincoln?”, in context of a passage on the former US president, a correct answer is most likely looking for a description referring to his job/role/occupation whereas the answer to the question “What is humidity?” is a definition. We argue that question class as well as named entities, when present in a question, directly affect the intent of the question and need to be

---

1https://cogcomp.seas.upenn.edu/Data/QA/QC/definition.html
<table>
<thead>
<tr>
<th>Question Pair</th>
<th>QBLEU</th>
<th>BLEURT</th>
<th>QSTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>(What was the title of Bob Dylan’s first album?; What was Bob Dylan’s first album called?)</td>
<td>0.744</td>
<td>0.816</td>
<td>0.928</td>
</tr>
<tr>
<td>(What was the name of Vincent’s brother; Who was Vincent’s brother?)</td>
<td>0.473</td>
<td>0.667</td>
<td>0.874</td>
</tr>
<tr>
<td>(Where in Germany was the composer Beethoven born?; Which city in Germany is the place of birth of Beethoven?)</td>
<td>0.278</td>
<td>0.709</td>
<td>0.548</td>
</tr>
<tr>
<td>(Who was Columbus?; Where is Columbus?)</td>
<td>0.508</td>
<td>0.671</td>
<td>0.0</td>
</tr>
<tr>
<td>(What was the name of Vincent’s brother?; What was the name of Vincent’s painting?)</td>
<td>0.832</td>
<td>0.519</td>
<td>0.0</td>
</tr>
<tr>
<td>(When did Freddie Mercury die?; How did Freddie Mercury die?)</td>
<td>0.779</td>
<td>0.799</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 1: Illustrative question pairs are shown with system-level scores for BLEU-1 (B1) and BLEU-4 (B4), METEOR, and ROUGE and as pair-level scores for QBLEU (Nema and Khapra, 2018), BLEURT (Sellam et al., 2020), and QSTS handled differently from other words in a question.

The metrics currently in use are based on word overlap and do not capture the semantics of questions as can be seen in the representative examples of similar and dissimilar questions in Table 1. The system-level scores of the traditional metrics (BLEU, METEOR, ROUGE), are shown in this table along with QBLEU values (the extension of BLEU scores for questions proposed by Nema, et al (2018)), as well as BLEURT scores that measure semantic similarity between two texts using BERT (Sellam et al., 2020; Devlin et al., 2019).

In Table 1, we note that none of the existing metrics are able to accurately assess the similarity or difference between the given question pairs and instead tend to assign high scores to dissimilar questions and low scores to simple rewritings of the questions with the same intent. In the rightmost column of Table 1, we show the values of our proposed **Question-Sensitive Text Similarity (QSTS)** scores assigned to these question pairs that are more representative. We discuss the design of QSTS in the rest of this paper. Our contributions are as follows:

1. We propose QSTS, a **Question-Sensitive Text Similarity** measure for comparing questions. Unlike existing measures, QSTS explicitly represents the question class and named entities present in a given question pair and combines them with dependency tree information and word embeddings to provide a more representative and interpretable measure of the semantic similarity between the two questions.

2. We evaluate QSTS on publicly-available datasets of similar questions available for QG/QA research. Our experiments indicate that our proposed measure provides a more accurate representation of question similarity compared to traditional measures employed for characterizing QG model performance.

3. We present the potential use of QSTS in reference-free evaluation for QG. The QSTS metric is able to reasonably characterize question quality of model-generated questions using passages that were used for generating them. This capability is representative of the human ability to judge whether a given question is fluent and relevant in the context of a given passage unlike existing measures that need reference questions for evaluation.

We demonstrate reference-free evaluation for QG using QSTS on a novel dataset, SimQG. SimQG contains human judgements for machine-generated questions from latest QG models for a selection of about 500 (reference question, passage) pairs from SQuAD (Rajpurkar et al., 2016). SimQG and an implementation of QSTS in Python have been made available for academic research.²

**Organization:** We present the details of comput-

² https://github.com/NUS-IDS/coling22_QSTS
ing QSTS in Section 2. Our novel dataset SimQG is described in Section 3. Experiments and results are described in Section 4 while closely-related work is summarized in Section 5. Finally, we conclude the paper with a summary and remarks on future directions in Section 6.

2 Question-Sensitive Text Similarity

A necessary aspect to capture while comparing two questions is a measure of whether the target intent behind the two questions is the same. As highlighted in the examples from Table 1, measuring simple lexical overlap between n-grams of two questions is insufficient for this purpose. In comparison, word and sentence representations (Pennington et al., 2014; Peters et al., 2018) are known to capture similarity between words despite the lexical mismatch. Extending this idea further, metrics based on contextual representations were developed for measuring similarity for text generation tasks such as translation and image captioning (Zhang et al., 2020; Sellam et al., 2020).

However, note that simple changes to words has significant changes in question meanings (“Who was Columbus” vs. “Where was Columbus?”) and embedding spaces learnt purely from word co-occurrence and contextual information from large corpora suffer from the drawback of overestimating scores to word pairs representing entities as well as question cues. Consequently, these measures tend to overestimate similarity in case of non-similar questions as shown in the last three examples in Table 1. We address the above issues by modeling three different question-specific aspects in QSTS:

**Question Class (QC)** or Answer-Type for a question refers to the constraints the question imposes on the “sought after answer” (Li and Roth, 2002). Li and Roth (2002) designed a two-level question class taxonomy (Footnote 1) for representing questions in TREC question answering tasks\(^4\) where the answers to questions can be assigned one of six coarse classes namely, Abbreviation, Entity, Description, Human, Location, and Numeric value. These six classes are further organized into 50 fine classes for a more specific classification of the answer type. For example, the coarse class “Human” includes fine classes for an individual, a group of individuals, a description of an individual, as well as the title assigned to an individual. Question class information has also been used to improve question answering and question generation performance (Tayyar Madabushi et al., 2018; Zhou et al., 2019).

We use question class information in QSTS to characterize if the two questions under consideration are seeking the same answer type. With an accurate question-class classifier, both questions in the first row of Table 1 are assigned the same question class in the QC taxonomy (referring to “creative pieces and inventions”) whereas the two questions in the fourth row are assigned classes corresponding to “description of an individual” versus “location” automatically capturing their different semantics. We can directly measure the question class similarity (\(qcsim\)) using the \(\delta\) function, where \(\delta_{ij} = 1\) if \(qc(q_i) = qc(q_j)\) and 0 otherwise where \(qc\) stands for the question class for question \(q_i\). To incorporate the hierarchical nature of the QC taxonomy, we modify this function to assign partial score of 0.5 if the coarse class matches for the two questions and 0.75 if one of question fine classes involves the catch-all “other” class. For example, the question classes assigned to the two questions in the third row correspond to “Location:Other” and “Location:city”, respectively.

**Named-Entities** when present in a question constrain the question with reference to the mentioned entity. For instance, if “Columbus” in the question “When was Columbus born?” is replaced by another name, it will become a completely different question. Therefore, similar to question classes, named entities require a hard measurement. To account for multi-word names and partial matches, we isolate the tokens in a given reference question referring to named entities and look for their presence in the generated question.\(^5\)

The named-entity similarity (\(nesim\)) is measured as the fraction of named-entity tokens in the reference question that are present in the generated question. That is, for a given reference question, “Who was Abraham Lincoln?” and the question “Who was Lincoln?”, the named-entity similarity score is computed as \(\frac{1}{2}\).

**Semantic Similarity** forms the third component of QSTS. We use the dependency parse of ques-

---

\(^1\)For instance, based on GLoVe embeddings (Pennington et al., 2014), (“Lincoln”, “Columbus”), and (“who”, “when”) have cosine similarity values of 0.659 and 0.608, respectively.

\(^2\)https://trec.nist.gov/data/qa.html

\(^3\)We use proper nouns in parts-of-speech tags and named-entity tags to identify name tokens.
tions to compute the semantic similarity between them. Dependency trees of sentences capture syntactic dependencies among the words in a sentence such as subject-object, modifier, and clausal links. Dependency-tree based kernels are widely-used in measuring sentence similarity (Croce et al., 2011; Özateş et al., 2016).

Let \( e=(h, \text{rel}, t) \) represent a directed edge in the dependency tree of a given question where the typed relation \( \text{rel} \) exists between two tokens, \( h \) and \( t \). Given the dependency edges of two questions (reference and generated), we match the dependency edges of the reference question \( (E(q_r)) \) with those of the generated question \( (E(q_g)) \) and pick the best matching or the most similar edge \( \forall e \in E(q_r) \). The edge similarity \( \delta_{\text{sim}}(e_m, e_g) \) is computed as

\[
\delta_{\text{sim}}(e_m, e_g) = \frac{\text{sim}(h_m, h_g) + \text{sim}(t_m, t_g)}{2}
\]

In the above formulation, \( \delta_{\text{sim}}(e_m, e_g) \) refers to the Kronecker \( \delta \) function that assigns a value of 1 if the two relation types are the same and zero otherwise and \( \text{sim}(a, b) \) is computed using cosine similarity of the word embeddings if \( a, b \) are non-name tokens. However, if either the head or tail of the edge is a name token, we use exact match on that side of the edge. That is, for the two edges \( e_m=(h_m, \text{rel}_m, t_m) \) and \( e_g=(h_g, \text{rel}_g, t_g) \), if \( h_m \) is a name token, \( \text{sim}(e_m, e_g) = \delta_{\text{sim}}(h_m, h_g) \text{sim}(t_m, t_g) \)

The same principle applies if \( t_m \) is a name token. The above formulation ensures that name tokens are not treated like regular tokens and a hard match is enforced while at the same time the edge similarity values stay in the range \([0, 1]\).

The final semantic similarity (\( \text{semsim} \)) of the two questions is the average similarity of the edges in the reference question that match best with the edges in the generated question. Since named entity tokens and question cue words are handled separately, only edges involving content words are considered in this computation. Additionally, we ignore edges representing less informative dependency relations such as “punctuation”, “possessive modifier” and seven others in line with previous works (Özateş et al., 2016).

**QSTS:** Note that each of the similarity functions, \( \text{qcsim}, \text{nesim}, \text{semsim} \) assign normalized scores between the two questions. These three scores can be summarized using the geometric mean (Fleming and Wallace, 1986) to obtain a single score between \([0, 1]\) for Question-Sensitive Text Similarity as

\[
\text{QSTS}(r, g) = (\text{qcsim}_{rg} \ast \text{nesim}_{rg} \ast \text{semsim}_{rg})^{1/3}
\]

The \( \text{QSTS} \) score is directional, the \( \text{nesim} \) and \( \text{semsim} \) computations are with respect to a given reference question. That is, for \( \text{nesim} \), we compute how many of the name tokens in a reference question are seen in the given/generated question and in \( \text{semsim} \), we compute the best matching edges from \( E(q_g), \forall e \in E(q_r) \). Note that this directional nature enables the computation of these two scores for (question, passage) pairs as well. To estimate if a question is valid for a passage, we can check if the named entities (when present) in the question can be found in the passage and if the dependency edges of the question are also supported in the passage. In this manner, \( \text{QSTS} \) provides for a reference-free evaluation of a question, given a passage context.

### 3 The SimQG dataset

Current models for QG are evaluated using QA datasets containing (passage, reference-question) pairs. Model-generated questions are compared against these reference questions using traditional metrics. It is our contention that given a passage context and questions generated by QG models against that context, several valid questions may be possible and it may not be representative to only compare generated questions against a specific given reference. To demonstrate this claim, we collected a novel dataset containing human judgments of machine-generated questions against their associated passage contexts and the reference questions available for these contexts.

Our novel dataset is based on SQuAD (Rajpurkar et al., 2016), a widely-used dataset in both QA and QG studies. About 500 (question, passage) pairs were randomly sampled from the test portion of the SQuAD dataset (used in (Zhou et al., 2018)). Recent QG models from ProphetNet (Qi et al., 2020), T5 (Raffel et al., 2020), and one of the early neural models based on Gated Self-Attention (GSA) networks (Zhao et al., 2018) were used for obtaining machine-generated questions. By choosing machine-generated questions from models with QG performance ranging from high (ProphetNet)
to significantly low (GSA), we seek to include questions in our dataset with varying degrees of answerability, fluency and relevance (Pan et al., 2020; Wang et al., 2020a).

Our annotation task on the crowdsourcing platform Amazon Mechanical Turk (AMT) was set up along the lines of previous QG works (Pan et al., 2020; Wang et al., 2020a). Each passage along with the machine generated question was examined by three independent crowdworkers to characterize if the question is (1) **Fluent**: Is the question grammatically correct, natural sounding, and semantically valid for the given passage context (yes=1.0, acceptable=0.5, no=0.0)?; (2) **Answerable**: Is the answer to the generated question present in the passage (yes=1.0, no=1.0)?; (3) **Relevant**: is the question relevant to the passage and only based on the content in the passage (yes=1.0, no=0.0)?

The workers were also asked to compare the machine-generated question with the reference question provided in SQuAD and to identify whether the question is similar to the reference question (score=1.0), similar but has less/more information compared to the reference question (score=0.5), different but has the same answer as the reference question (score=0), or related but different (score=0) and finally very different from the reference question (score=0).

By averaging worker scores for each question, we obtain relevance/fluency/answerability scores for each (machine-generated question, passage) pair as well as a similarity score for pairs of (machine-generated, reference) question pairs all in the range $[0,1]$ and by suitably thresholding at 0.5, we can obtain pairs of similar and dissimilar questions for our study as well as questions that are not fluent, not relevant, not answerable, and not similar to the given reference question.

A summary of **SimQG** is provided in Table 2. As seen in this table, all three QG models generate reasonably fluent questions. In accordance with the published QG performance numbers of these models on SQuAD dataset, the number of non-fluent and non-answerable questions is the highest for GSA, lowest for ProphetNet (PrptNet) and in-between for the T5-based model (Qi et al., 2020; Zhao et al., 2018). In all three models, the number of non-relevant machine-generated questions is very low (2-6%) whereas the machine-generated question was considered not similar to the reference question in 40-50% of the cases. We posit that this high disparity is indicative of why QG models need evaluation measures that are not based only on reference questions.

<table>
<thead>
<tr>
<th>QGModel</th>
<th>!Flu</th>
<th>!Rel</th>
<th>!Ans</th>
<th>!Sim</th>
</tr>
</thead>
<tbody>
<tr>
<td>PrptNet (300)</td>
<td>0.66%</td>
<td>2%</td>
<td>4.66%</td>
<td>45%</td>
</tr>
<tr>
<td>GSA (100)</td>
<td>7%</td>
<td>4%</td>
<td>19%</td>
<td>47%</td>
</tr>
<tr>
<td>T5 (100)</td>
<td>4%</td>
<td>6%</td>
<td>12%</td>
<td>38%</td>
</tr>
<tr>
<td>All(500)</td>
<td>2.6%</td>
<td>3.2%</td>
<td>9%</td>
<td>44%</td>
</tr>
</tbody>
</table>

Table 2: Summary of **SimQG** dataset. #Qs is the number of questions whereas !Flu, !Rel, !Ans, !Sim columns refer to the percentages of questions that are not fluent, not relevant, not answerable, and not similar to the given reference question.

Additional notes on data collection: On the AMT platform, we required the crowdworkers to have greater than 95% HIT approval rate, a minimum of 10,000 HITs, and be located in the United States and clear a qualification test to be able to work on our task. Each worker was paid $0.30 per HIT. We met the **ethics, quality, and reliability** considerations for our collected dataset as follows: As part of the AMT data collection process, the **anonymity and privacy** of the crow workers is already ensured. Furthermore, the settings for the HIT approval rates, and location of the worker, described previously are set similar to previous QA/QG data collection efforts to ensure the English language skills of the data annotators and thus the **quality** of the collected dataset. A total of 7 workers helped in creating our dataset. About 47% of the workers who attempted the qualification test were able to obtain a score of 80% or more and gained the eligibility to work on our task. Their annotations can, therefore, be considered reasonably reliable on average.

4 Experiments

Baseline Measures: We demonstrate the performance of our proposed QSTS measure by comparing with several existing measures. The first set of measures are traditionally employed in various text generation tasks including QG and comprise of BLEU, METEOR, and ROUGE scores (Papineni et al., 2002; Lavie and Agarwal, 2007; Lin, 2004). All these measures are based on n-gram overlap between the generated text and the reference text (of the same “type”, for example, two summaries, or two sentences).
The QBLEU metric was designed specifically for QG systems and includes the notion of answerability, that is, does the question include enough information to enable answer retrieval for the given question (Nema and Khptra, 2018)? To this end, various weights are estimated and incorporated for question cue words, content words, named entities and combined linearly to assign answerability score for a question. Answerability is further combined with the traditional BLEU score to obtain a QBLEU score.\footnote{https://github.com/PrekshaNema25/Answerability-Metric}

A recent research direction involves the use of transformers for learning metrics for text generation tasks (Zhang et al., 2020; Sellam et al., 2020). Based on its state-of-the-art performance on various BERT tasks compared to other variants such as BLEURT, we include BLEURT as one of our baselines for comparing questions. To the best of our knowledge, BLEURT has not been specifically evaluated for matching questions and we seek to bridge this gap as part of our experiments.\footnote{https://github.com/google-research/bleurt}

Datasets: We used two existing QA/QG datasets with paraphrase information for evaluation. The first is the ComQA dataset that includes about 3.3K paraphrase pairs (Abujabal et al., 2019) while the second is the recently-compiled FIRS dataset, containing approximately 5K question pairs (Deschamps et al., 2021). The FIRS dataset includes rewrites of a given question which also include an extra fact from a knowledge base. That is, the rewritten question has the same intent as the original question but includes additional facts of relevant named entities. We randomly selected one of questions from each paraphrase clusters in ComQA as the reference question whereas in FIRS, the original question forms the reference question.

In addition, we evaluate on questions from the Quora Question Pairs (QQP) dataset.\footnote{https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs} QQP is a large dataset of about 400k question pairs obtained from Quora and includes duplicate and non-duplicate labels indicative of whether the intent of the two questions is the same. Note that this dataset is not used for QG since passage contexts and answers are unavailable. Moreover, the labels are known to be noisy in this dataset, and the questions on community forums are stylistically different from standard QG (António Rodrigues et al., 2017). Despite these differences, we study a randomly selected 5% sample of the QQP dataset separated into duplicate pairs (QQP-Dup) and non-duplicate pairs (QQP-ND). Finally, we provide evaluation on SimQG, the dataset specifically collected by us to model QG contexts (Section 3).

Question Class Identification: We trained our question class classifier on the widely-used TREC dataset (Li and Roth, 2002). A T5-large model\footnote{https://huggingface.co/t5-large} fine-tuned for this task obtains a test performance on par with state-of-the-art results with a classification accuracy of 92% on the fine-level classes (50 classes) and an accuracy of 97% on the six coarse classes (Reimers and Gurevych, 2019). When computing QSTS scores for question pairs where the question classes cannot be assumed to be the same (such as QQP-ND and SimQG), predictions with this model were used.

Other Settings: For computing QSTS scores, we need the dependency parse, parts-of-speech and named-entity tags for questions. We used the Stanza library for this purpose.\footnote{https://stanfordnlp.github.io/stanza/} Since the name-tokens and question class information are treated separately, we avoid contextual and sentence-level embeddings that are time-consuming to estimate (Peters et al., 2018; Reimers and Gurevych, 2019) and instead directly use word embeddings from GloVe that only involves lookup (Pennington et al., 2014). All QG and QC experiments, and metrics that involve deep learning models were performed on a single GPU of an Nvidia Tesla cluster and take time between 1-12 hours based on the experiment setting and dataset size. The code for QSTS and the SimQG dataset have been released for academic research.

4.1 Results and Observations

Comparison of Measures: We compare QSTS against existing baseline measures on similar questions from ComQA and FIRS, as well as duplicate and non-duplicate question datasets QQP-Dup and QQP-ND, respectively. An ideal measure should assign high scores (close to 1) to similar questions and low scores (close to 0) to dissimilar ones. We show the system-level BLEU, METEOR, and ROUGE score as well as average and standard deviation of QBLEU, BLEURT, and QSTS scores
in Table 3.

**Performance on Paraphrase datasets:** We see in Table 3 that QSTS is significantly better than other measures on FIRS and QQP-Dup datasets, and is on par with the BLEURT measure on ComQA. We analyzed ComQA further to gain insight into where QSTS breaks down. From Equation 1, the QSTS score is zero when any of the component scores, qcsim, nesim, and semsim, is zero. That is, when the question classes, named-entities, or the content words of the two questions do not match. In ComQA, the QSTS score was zero for 16.9% question pairs with the qcsim, nesim, and semsim scores being independently zero in 5%, 8.6%, and 4.6% of the question pairs, respectively.

Since question paraphrases should, ideally, have the same question class but qcsim is zero for 5% of the cases, we can attribute these mismatches to the errors made by the question class predictor. However, we also note that this could be caused by erroneous pairs present in ComQA such as (“when did Judy Garland first marry?”; “who was Judy Garlands first married to?”) where the question classes are indeed different and were predicted correctly as “NUM:date” and “HUM:ind” by our question class predictor.

Furthermore, ComQA also has instances where mentions of the same named-entity have typos and other differences. For example, pairs such as (“what is muhamad alis real name?”; “what is mahummad ali birth name?”) and (“who was the german fascist leader during world war 2?”; “what man was the leader of germany during ww2?”).

Finally, about 8% of the questions in ComQA appear to be in search-engine style (“the first american in outer space?”) and do not have any of the question cue words. Noisy inputs affect the type of dependency edges between content words and may result in zero semsim scores. Overall, QSTS is not fully-equipped to handle noisy paraphrases since errors in the component scores are penalized severely (Equation 1).

**Performance on Non-Paraphrases:** On the QQP-ND dataset containing non-duplicate question pairs, simple n-gram based measures seem to do better than embedding-based BLEURT and QSTS measures which overestimate the similarity scores. As mentioned in Section 2, the QSTS scores are directional and also depend on the predictions from the question class classifier. Therefore, in the given non-duplicate pair from QQP-ND, (“How does one become an angel investor?”; “How do I get a job at Goldman Sachs?”), the QSTS scores change from 0 to 0.718 depending on which question forms the “reference”. Moreover, QC predictions may not always be accurate considering the stylistic differences in QQP questions when compared to those from TREC. Questions in QQP include conjunctions such as “How do you bake pork chops in an oven and how long should you bake them?” and multi-sentence questions such as “I have completed my MBA with . . . in PSU. I want to work abroad, how do I start?”.

We note that accurate measurement of dissimilar questions is not a big concern in QG contexts. For a given passage context, a good QG model is unlikely to generate a question comprising of completely arbitrary words in contrast with some pairs in QQP-ND such as (“How can I get perfect idea about best golf carts?”; “Is it hard to get a job in US after MIS without prior work experience?”).

**Reference-free Evaluation on SimQG:** Using the human-assigned scores for similarity between machine-generated and reference question pairs in SimQG, we threshold at 0.5 to obtain pairs of questions considered similar and for these questions, we compute QSTS scores between the machine-generated questions and the corresponding passages. In other words, how many machine-generated questions can we correctly assign a score value ≥ 0.5 when the passage is used instead of comparing with the reference?

Similarly, for the set of machine-generated questions judged as non-fluent, non-relevant, and non-answerable by humans, how many questions are correctly assigned scores < 0.5 based on the pas-

<table>
<thead>
<tr>
<th>Dataset</th>
<th>B1</th>
<th>B4</th>
<th>METEOR</th>
<th>ROUGE</th>
<th>QBLEU</th>
<th>BLEURT</th>
<th>QSTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ComQA</td>
<td>0.602</td>
<td>0.287</td>
<td>0.373</td>
<td>0.566</td>
<td>0.594±0.151</td>
<td>0.696±0.110</td>
<td>0.692±0.298</td>
</tr>
<tr>
<td>FIRS</td>
<td>0.557</td>
<td>0.430</td>
<td>0.485</td>
<td>0.695</td>
<td>0.554±0.187</td>
<td>0.728±0.110</td>
<td>0.866±0.231</td>
</tr>
<tr>
<td>QQP-Dup</td>
<td>0.561</td>
<td>0.277</td>
<td>0.334</td>
<td>0.545</td>
<td>0.449±0.231</td>
<td>0.711±0.112</td>
<td>0.754±0.283</td>
</tr>
<tr>
<td>QQP-ND</td>
<td>0.342</td>
<td>0.158</td>
<td>0.204</td>
<td>0.344</td>
<td>0.289±0.252</td>
<td>0.491±0.169</td>
<td>0.388±0.391</td>
</tr>
</tbody>
</table>

Table 3: Question Similarity Metrics Evaluated on Existing Datasets

\[ \text{QSTS} = \begin{cases} 0 & \text{if any of } qcsim, nesim, \text{ or semsim is zero} \\ \frac{\text{correctly assigned scores}}{\text{total number of question pairs}} & \text{otherwise} \end{cases} \]
sage. The results of these computations are illustrated in Table 4.

<table>
<thead>
<tr>
<th>Setting</th>
<th>#Qs</th>
<th>NoQC</th>
<th>withQC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similar</td>
<td>280</td>
<td>66.1%</td>
<td>56.4%</td>
</tr>
<tr>
<td>Non-Fluent</td>
<td>13</td>
<td>30.7%</td>
<td>76.9%</td>
</tr>
<tr>
<td>Non-Relevant</td>
<td>16</td>
<td>56.3%</td>
<td>81.3%</td>
</tr>
<tr>
<td>Non-Answerable</td>
<td>45</td>
<td>40.0%</td>
<td>66.6%</td>
</tr>
</tbody>
</table>

Table 4: Percentages of similar, non-fluent, non-relevant and non-answerable questions correctly identified in SimQG by QSTS in the reference-free setting. #Qs refers to the number of questions whereas NoQC and withQC refer to with and without the question class information, respectively.

The percentage of questions correctly scored using reference-free QSTS is about 56% when question class information obtained from reference questions is incorporated (“withQC” column in Table 4). The percentage is, however, significantly higher (66%) when question class information is not considered (“NoQC”). This difference suggests that valid questions are being generated for a given passage context despite having different question classes. However, when question class information is incorporated into QSTS computation, we are able to determine non-fluent, non-relevant, and non-answerable questions with significantly higher accuracy as observed in Table 4.

In practice, it may not be unreasonable to assume that the expected question class is known a priori, considering the current state-of-the-art QG performance is obtained in the answer-aware (as opposed to answer agnostic) setting when the answer span is assumed to be known and used as a signal while learning QG (Pan et al., 2019). Even without explicit question class information, QSTS can correctly identify relevant and non-relevant questions with reasonable accuracies. Given that this is the first method to do so without a known reference question, this is an exciting result.

In contrast, the traditional measures as well as QBLEU and BLEURT expect similar types of texts for their computation. In our experiments, when QBLEU and BLEURT measures are computed using generated questions and passages as inputs, both measures were unable to correctly assign scores >= 0.5 to any of the similar questions. That is, the percentage correct values in the top row of Table 4 are zeros for both these measures. Anecdotal examples of question, passage pairs scored with our QSTS measure are provided in Table 5 for illustration. QSTS correctly assigns high scores (indicating relevant) to the top-two (question, passage) pairs and lower scores (less than 0.5 indicating non-relevant) to the bottom two pairs.

Finally, the average QSTS scores for the test split of SQuAD (Zhou et al., 2018) with ProphetNet (Qi et al., 2020), T5 (Footnote 6), and GSA (Zhao et al., 2018) models are shown below.

<table>
<thead>
<tr>
<th>Model</th>
<th>QSTS</th>
<th>BLEU-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ProphetNet</td>
<td>0.506 (± 0.386)</td>
<td>25.80</td>
</tr>
<tr>
<td>T5</td>
<td>0.407 (± 0.376)</td>
<td>21.32</td>
</tr>
<tr>
<td>GSA</td>
<td>0.344 (± 0.370)</td>
<td>16.38</td>
</tr>
</tbody>
</table>

The BLEU-4 scores published for these models are shown in the rightmost column of the table and though these published numbers use different data splits for SQuAD compared to ours, we would like to highlight that the overall performance trend of these models as seen by their BLEU-4 scores is also captured by QSTS.

In summary, QSTS presents as a viable alternative to traditional measures for evaluating QG in terms of its interpretable score components. Moreover, QSTS enables a reference-free evaluation for real-world QG scenarios where precompiled lists of reference questions are unavailable.

**Limitations:** Although QSTS addresses several problems with existing QG metrics (Table 1), we note the following caveats that need further work.

1. The QSTS function is sensitive to the component scores. Though geometric mean is suggested for summarizing normalized scores (Fleming and Wallace, 1986), and yields higher performance compared to other mean functions in our experiments, other combining functions can be investigated in future.

2. New question-type ontologies are being developed to cover contexts different from extractive QA such as questions within dialog (Cao and Wang, 2021; Svikhnushina et al., 2022; Malhotra et al., 2022). High-accuracy question-class predictors need to be trained for using QSTS in these contexts.

3. None of the existing metrics as well as our proposed measure directly incorporate notions such as fluency, interesting-ness, and answerability that humans are able to assess naturally.
5 Related Work

Models for question generation are being rapidly developed in current NLP research. We refer our readers to a survey article by Pan, et al (2019) for an overview on challenges, existing approaches, and applications for this task. Similar to the standard practice in NLG tasks, question generation has been evaluated using n-gram overlap based metrics such as BLEU (Papineni et al., 2002), METEOR Lavie and Agarwal (2007) and ROUGE (Lin, 2004). While previous studies have found these metrics inadequate for tasks such as summarization, paraphrase generation, and translation (Callison-Burch et al., 2006; Shen et al., 2022), Nema, at al. (2018) specifically study their drawbacks in context of question generation models. Indeed, similar to our approach, they isolate various types of tokens in questions and assign tuned weights to question cue-words, content words, function words, and named entities to compute “answer-ability” for a question.

In parallel studies, the notion of unsupervised evaluation metrics were studied for dialog systems and machine translation (Liu et al., 2016; Fomicheva et al., 2020) while metric learning was explored for several NLG tasks using transformers in BERTscore, BBScore, and BLEURT (Zhang et al., 2020; Sellam et al., 2020; Shen et al., 2022).

We have highlighted cases where these existing metrics fall short for question comparison and specifically propose reference-free evaluation possibilities for question generation. Reference-free evaluation was previously studied for NLG tasks such as machine translation (Agrawal et al., 2021) and essay grading (Fomicheva et al., 2020).

6 Conclusions and Future Work

We discussed existing metrics for question generation evaluation and highlighted cases where a deeper understanding of question semantics need to be modeled by metrics for a more representative evaluation. As an alternative, we designed the question-sensitive text similarity metric (QSTS) that comprises of interpretable scoring components. We also underscored the need for reference-free evaluation in QG systems. The potential of QSTS in serving this purpose was demonstrated on a novel dataset SimQG, compiled from human judgments on (machine-generated question, passage) pairs. QSTS provides, to our knowledge, the first approach to characterizing QG system performance in practical deployments where reference questions may not always be available.

As can be seen in experiments, there is still a large room for improvement for question similarity computation as well as QG evaluation. In future, we hope to pursue these directions further as well as study metric learning approaches for a reference-free evaluation.

Acknowledgments

This research/project is supported by the National Research Foundation, Singapore under its Industry Alignment Fund – Pre-positioning (IAF-PP) Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. We thank our anonymous reviewers for their insightful comments. Content related to the error analysis on ComQA and limitations was added as a result of these reviews.

Ethics Statement

This research was conducted in conformance with the ACM Code of Ethics.
References


Qingbao Huang, Mingyi Fu, Linzhang Mo, Yi Cai, Jingyun Xu, Pijian Li, Qing Li, and Ho-fung Leung. 2021. Entity guided question generation with contextual structure and sequence information capturing. Proceedings of the AAAI Conference on Artificial Intelligence, 35(14).


Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic graphs for generating deep questions. In ACL.


Jian Wang, Junhao Liu, Wei Bi, Xiaojiang Liu, Kejing He, Ruifeng Xu, and Min Yang. 2020a. Improving knowledge-aware dialogue generation via knowledge base question answering. AAAI.


