How to Ask Better Questions? A Large-Scale Multi-Domain Dataset for Rewriting Ill-Formed Questions

Zewei Chu,[1]* Mingda Chen,[2]*† Jing Chen,[3]† Miaosen Wang,[3]†
Kevin Gimpel,[2] Manaal Faruqui,[3] Xiance Si[3]
[1]The University of Chicago, 5730 S Ellis Ave, Chicago, IL 60637, USA
[2]Toyota Technological Institute at Chicago, 6045 S Kenwood Ave, Chicago, IL 60637, USA
[3]Google Assistant, 1600 Amphitheatre Pkwy, Mountain View, CA 94043, USA
zeweichu@uchicago.edu, {mchen, kgimpel}@ttic.edu
{chenjin, miaosen, mfaruqui, sxc}@google.com

Abstract

We present a large-scale dataset for the task of rewriting an ill-formed natural language question to a well-formed one. Our multi-domain question rewriting (MQR) dataset is constructed from human contributed Stack Exchange question edit histories. The dataset contains 427,719 question pairs which come from 303 domains. We provide human annotations for a subset of the dataset as a quality estimate. When moving from ill-formed to well-formed questions, the question quality improves by an average of 45 points across three aspects. We train sequence-to-sequence neural models on the constructed dataset and obtain an improvement of 13.2% in BLEU-4 over baseline methods built from other data resources. We release the MQR dataset to encourage research on the problem of question rewriting.[1]

Introduction

Understanding text and voice questions from users is a difficult task as it involves dealing with "word salad" and ill-formed text. Ill-formed questions may arise from imperfect speech recognition systems, search engines, dialogue histories, inputs from low bandwidth devices such as mobile phones, or second language learners, among other sources. However, most downstream applications involving questions, such as question answering and semantic parsing, are trained on well-formed natural language. In this work, we focus on rewriting textual ill-formed questions, which could improve the performance of such downstream applications.

Faruqui and Das (2018) introduced the task of identifying well-formed natural language questions. In this paper, we take a step further to investigate methods to rewrite ill-formed questions into well-formed ones without changing their semantics. We create a multi-domain question rewriting dataset (MQR) from human contributed Stack Exchange question edit histories.[2] This dataset provides pairs of questions: the original ill-formed question and a well-formed question rewritten by the author or community contributors. The dataset contains 427,719 question pairs which come from 303 domains. The MQR dataset is further split into TRAIN and DEV/TEST, where question pairs in DEV/TEST have less n-gram overlap but better semantic preservation after rewriting. Table 1 shows some example question pairs from the MQR DEV split.

Our dataset enables us to train models directly for the task of question rewriting. We train neural generation models on our dataset, including Long-Short Term Memory networks (LSTM; Hochreiter and Schmidhuber 1997) with attention (Luong, Pham, and Manning 2015) and transformers (Vaswani et al. 2017). We show that these models consistently improve the well-formedness of questions although sometimes at the expense of semantic drift. We compare to approaches that do not use our training dataset, including general-purpose sentence paraphrasing, grammatical error correction (GEC) systems, and round trip neural machine translation. Methods trained on our dataset greatly outperform those developed from other resources. Augmenting our training set with additional question pairs such as Quora or Paralex question pairs (Fader, Zettlemoyer, and Etzioni 2013) has mixed impact on this task. Our findings from the benchmarked methods suggest potential research directions to improve question quality.

To summarize our contributions:

- We propose the task of question rewriting: converting textual ill-formed questions to well-formed ones while preserving their semantics.
- We construct a large-scale multi-domain question rewriting dataset MQR from human generated Stack Exchange question edit histories. The development and test sets are of high quality according to human annotation. The training set is of large-scale. We release the MQR dataset to encourage research on the question rewriting task.
- We benchmark a variety of neural models trained on the MQR dataset, neural models trained with other question rewriting datasets, and other paraphrasing techniques. We find that models trained on the MQR and Quora datasets combined followed by grammatical error correction perform the best in the MQR question rewriting task.

*Work performed during internship at Google.

†Contributed equally.

[1] https://github.com/ZeweiChu/MQR

[2] https://archive.org/download/stackexchange

Ill-formed	Well-formed	Category
Spaghetti carbonara, mixing	How to mix a spaghetti carbonara?	cooking
Ethical Investing... where to begin?	How to begin ethical investing?	money
charging canon sx 700 battery through powerbank	Can I charge a Canon SX 700 battery using a mobile powerbank?	photo
H1B Visa consulate interview timeline	What is the timeline for an H1B visa consulate interview?	expatriates
Hanging weight from drywall ceiling	How much weight can I hang from a drywall ceiling?	diy

Table 1: Examples of pairs of ill-formed and well-formed questions from the MQR dataset.

Related Work

Query and Question Rewriting

Methods have been developed to reformulate or expand search queries (Jones et al. 2006). Sometimes query rewriting is performed for sponsored search (Zhang et al. 2007; Zhang and Jones 2007). This work differs from our goal as we rewrite ill-formed questions to be well-formed.

Some work rewrites queries by searching through a database of query logs to find a semantically similar query to replace the original query. De Bona et al. (2010) compute query similarities for query ranking based on user click information. Dong et al. (2017) learn paraphrases of questions to improve question answering systems. Kumar, Dandapat, and Chordia (2018) translate queries from search engines into natural language questions. They used Bing's search logs and their corresponding clicked question page as a query-to-question dataset. We work on question rewriting without any database of question logs.

Actively rewriting questions with reinforcement learning has been shown to improve QA systems (Buck et al. 2018). This work proposes to rewrite questions to fulfill more general quality criteria.

Paraphrase Generation

A variety of paraphrase generation techniques have been proposed and studied (Barzilay and Lee 2003; Bannard and Callison-Burch 2005; Androutsopoulos and Malakasiotis 2010; Madnani and Dorr 2010; Malakasiotis and Androutsopoulos 2011; Li et al. 2019). Recently, Gupta et al. (2018) use a variational autoencoder to generate paraphrases from sentences and Li et al. (2018) use deep reinforcement learning to generate paraphrases. Several have generated paraphrases by separately modeling syntax and semantics (Iyyer et al. 2018; Chen et al. 2019).

Paraphrase generation has been used in several applications. Cho, Xie, and Campbell (2019) use paraphrase generation as a data augmentation technique for natural language understanding. Iyyer et al. (2018) and Ribeiro, Singh, and Guestrin (2018) generate adversarial paraphrases with surface form variations to measure and improve model robustness. Wieting and Gimpel (2018) generate paraphrases using machine translation on parallel text and use the resulting sentential paraphrase pairs to learn sentence embeddings for semantic textual similarity.

Our work focuses on question rewriting to improve question qualities, which is different from general sentence paraphrasing.

Text Normalization

Text normalization (Sproat et al. 2001) is the task of converting non-canonical language to "standard" writing. Non-canonical language frequently appears in informal domains such as social media postings or other conversational text, user-generated content, such as search queries or product reviews, speech transcriptions, and low-bandwidth input settings such as those found with mobile devices. Text normalization is difficult to define precisely and therefore difficult to provide gold standard annotations and evaluate systems for (Eisenstein 2013). In our setting, rewriting questions is defined implicitly through the choices made by the Stack Exchange community with the goals of helpfulness, clarity, and utility.

Task Definition: Question Rewriting

Given a question q_i, potentially ill-formed, the question rewriting task is to convert it to a well-formed natural language question q_w while preserving its semantics and intention. Following Faruqui and Das (2018), we define a well-formed question as one satisfying the following constraints:

- The question is grammatically correct. Common grammatical errors include misuse of third person singular or verb tense.
- The question does not contain spelling errors. Spelling errors refer specifically to typos and other misspellings, but not to grammatical errors such as third person singular or tense misuse in verbs.
- The question is explicit. A well-formed question must be explicit and end with a question mark. A command or search query-like fragment is not well-formed.

MQR Dataset Construction and Analysis

We construct our Multi-Domain Question Rewriting (MQR) dataset from human contributed Stack Exchange question edit histories. Stack Exchange is a question answering platform where users post and answer questions as a community. Stack Exchange has its own standard of good questions,[3] and their standard aligns well with our definition of well-formed questions. If questions on Stack Exchange do not meet their quality standards, members of the community often volunteer to edit the questions. Such edits typically correct spelling and grammatical errors while making the question more explicit and easier to understand.

[3] https://meta.stackexchange.com/questions/92074/what-can-i-do-when-getting-this-question-body-does-not-meet-our-quality-standar

Question	Spelling	Grammar	Explicit	Remark
How to remove water-based paint?	1	1	1	
how can I make quark the music player to work in unity, natty?	0	0	1	
What is the value to checking in broken unit tests?	1	0	1	to → of
No room for RO drain saddle?	1	1	0	

Table 2: Examples given to annotators for binary question quality scores.

Question 1	Question 2	Equivalent
How to add a lightbox?	How to add a lightbox to class mix?	0
how to get md5sum of a string directly in terminal	How to get the MD5 hash of a string directly in the terminal?	1

Table 3: Example question pairs given to annotators to judge semantic equivalence.

We use 303 sub areas from Stack Exchange data dumps.[4] The full list of area names is in the appendix. We do not include Stack Overflow because it is too specific to programming related questions. We also exclude all questions under the following language sub areas: *Chinese*, *German*, *Spanish*, *Russian*, *Japanese*, *Korean*, *Latin*, *Ukrainian*. This ensures that the questions in MQR are mostly English sentences. Having questions from 303 Stack Exchange sites makes the MQR dataset cover a broad range of domains.

We use "PostHistory.xml" and "Posts.xml" tables of each Stack Exchange site data dump. If a question appears in both "PostHistory.xml" and "Posts.xml", it means the question was modified. We treat the most up-to-date Stack Exchange questions as a well formed-question and treat its version from "PostHistory.xml" as ill-formed. "PostHistory.xml" only keeps one edit for each question, so the MQR dataset does not contain duplicated questions.

The questions in the Stack Exchange raw data dumps do not always fulfill our data quality requirements. For example, some questions after rewriting are still not explicit. Sometimes rewriting introduces or deletes new information and cannot be done correctly without more context or the question description. We thus perform the following steps to filter the question pairs:

1. All well-formed questions in the pairs must start with "how", "why", "when", "what", "which", "who", "whose", "do", "where", "does", "is", "are", "must", "may", "need", "did", "was", "were", "can", "has", "have", "are". This step is performed to make sure the questions are explicit questions but not statements or commands.
2. To ensure there are no sentences written in non-English languages, we keep questions that contain 80% or more of valid English characters, including punctuation.[5]

This yields the MQR dataset. We use the following heuristic criteria to split MQR into TRAIN, DEV, and TEST sets:

1. The BLEU scores between well-formed and ill-formed questions (excluding punctuation) are lower than 0.3 in DEV and TEST to ensure large variations after rewriting.
2. The lists of verbs and nouns between well-formed and ill-formed questions have a Jaccard similarity greater than 0.8 in DEV and TEST. We split DEV and TEST randomly and equally. This yields 2,112 instances in DEV and 2,113 instances in TEST.
3. The rest of the question edit pairs (423,495 instances) are placed in the TRAIN set.

Examples are shown in Table 1. We release our TRAIN/DEV/TEST splits of the MQR dataset to encourage research in question rewriting.

Dataset Quality

To understand the quality of the question rewriting examples in the MQR dataset, we ask human annotators to judge the quality of the questions in the DEV and TEST splits (abbreviated as DEVTEST onward). Specifically, we take both ill-formed and well-formed questions in DEVTEST and ask human annotators to annotate the following three aspects regarding each question (Faruqui and Das 2018):

1. Is the question grammatically correct?
2. Is the spelling correct? Misuse of third person singular or past tense in verbs are considered grammatical errors instead of spelling errors. Missing question mark in the end of a question is also considered as spelling errors.
3. Is the question an explicit question, rather than a search query, a command, or a statement?

The annotators were asked to annotate each aspect with a binary (0/1) answer. Examples of questions provided to the annotators are in Table 2. We consider all "How to" questions ("How to unlock GT90 in Gran Turismo 2?") as grammatical. Although it is not a complete sentence, this kind of question is quite common in our dataset and therefore we choose to treat it as grammatically correct.

The ill-formed and well-formed questions are shuffled so the annotators do not have any prior knowledge or bias regarding these questions during annotation. We randomly sample 300 questions from the shuffled DEVTEST questions, among which 145 examples are well-formed and 155

[4] https://archive.org/download/stackexchange

[5] The list of valid characters after lowercasing is: 0123456789abcdefghijklmnopqrstuvwxyz . , / ? : ; ’ []_ + - = ! @ # $ % &̂ * ( ) | { } <> ‘ ” ’ ” and space

Quality						Semantic Equivalence
Ill-formed			Well-formed			
Spelling	Grammar	Explicit	Spelling	Grammar	Explicit	
139/310=0.45	166/310=0.54	175/310=0.56	282/290=0.97	271/290=0.93	286/290=0.99	183/200=0.92

Table 4: Summary of manual annotations for instances sampled from the DEV and TEST portions of the MQR dataset. "Quality" are the average quality scores, broken down into three aspects. "Semantic Equivalence" is the percentage of question pairs in which the ill-formed and well-formed questions are semantically equivalent. The scores are averages of binary scores across both annotators.

	TRAIN	DEVTEST
# Categories	303	166
Mean # instances per category	320.4	25.5
Std # instances per category	754.8	47.1
Min # instances per category	1	1
Max # instances per category	6237	295

Table 5: Statistics of question pairs ("instances") from Stack Exchange categories in the MQR dataset.

are ill-formed. Two annotators produce a judgment for each of the three aspects for all 300 questions.

The above annotation task considers a single question at a time. We also consider an annotation task related to the quality of a question *pair*, specifically whether the two questions in the pair are semantically equivalent. If rewriting introduces additional information, then the question rewriting task may require additional context to be performed, even for a human writer. This may happen when a user changes the question content or the question title is modified based on the additional description about the question. In the MQR dataset, we focus on question rewriting tasks that can be performed without extra information.

We randomly sample 100 question pairs from DEVTEST for annotation of semantic equivalence. Two annotators produced binary judgments for all 100 pairs. Example pairs are shown in Table 3.

Table 4 summarizes the human annotations of the quality of the DEVTEST portion of the MQR dataset. We summed up the binary scores from two annotators. There are clear differences between ill-formed and well-formed questions. Ill-formed question are indeed ill-formed and well-formed questions are generally of high quality. The average score over three aspects improves by 45 points from ill-formed to well-formed questions. Over 90% of the question pairs possess semantic equivalence, i.e., they do not introduce or delete information. Therefore, the vast majority of rewrites can be performed without extra information.

The Cohen's Kappa inter-rater reliability scores (McHugh 2012) are 0.83, 0.77, and 0.89 respectively for the question quality annotations, and 0.86 for question semantic equivalence. These values show good inter-rater agreement on the annotations of the qualities and semantic equivalences of the MQR question pairs.

Dataset Domains

As the MQR dataset is constructed from 303 sub areas of the Stack Exchange networks, it covers a wide range of question domains. Table 5 summarizes the number of categories in the TRAIN and DEVTEST portions of the MQR dataset, as well as the mean, standard deviation, minimum, and maximum number of instances per categories.

The number of questions from each sub area is not evenly distributed due to the fact that some sub areas are more popular and have more questions than the others, but the DEV/TEST splits still cover a reasonably large range of domains.

The most common categories in DEV and TEST are "diy"(295), "askubuntu"(288), "math"(250), "gaming"(189), and "physics"(140). The least common categories are mostly "Meta Stack Exchange" websites where people ask questions regarding the policies of posting questions on Stack Exchange sites. The most common categories in TRAIN are "askubuntu"(6237), "math"(5933), "gaming"(3938), "diy"(2791), and "2604"(scifi).

Models and Experiments

In this section, we describe the models and methods we benchmarked to perform the task of question rewriting.

To evaluate model performance, we apply our trained models to rewrite the ill-formed questions in TEST and treat the well-formed question in each pair as the reference sentence. We then compute BLEU-4 (Papineni et al. 2002), ROUGE-1, ROUGE-2, ROUGE-L (Lin 2004), and METEOR (Banerjee and Lavie 2005) scores.[6] As a baseline, we also evaluate the original ill-formed question using the automatic metrics.

Models Trained on MQR

Transformer. We use the Tensor2Tensor (Vaswani et al. 2018) implementation of the transformer model (Vaswani et al. 2017). We use their "transformer_base" hyperparameter setting. The details are as follows: batch size 4096, hidden size 512, 8 attention heads, 6 transformer encoder and decoder layers, learning rate 0.1 and 4000 warm-up steps. We train the model for 250,000 steps and perform early stopping using the loss values on the DEV set.

[6] The BLEU-4 and METEOR scores are calculated using https://github.com/Maluuba/nlg-eval. ROUGE-1, ROUGE-2, and ROUGE-L are calculated using https://github.com/pltrdy/rouge.

	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Ill-formed	5.9	50.9	19.4	45.5	33.4
Models trained on MQR					
LSTM seq-to-seq with attention	19.2	55.8	28.3	52.8	32.7
Transformer	**22.1**	**59.8**	**32.2**	**56.6**	**36.4**
Methods built from other resources					
Grammatical error correction	13.1	52.4	24.4	47.5	34.4
Round trip NMT (Pivot: De)	9.9	41.6	16.8	38.2	28.4
Round trip NMT (Pivot: Fr)	9.3	40.4	15.7	36.9	27.5
Paraphrase generator trained on ParaNMT	4.9	24.8	7.5	21.8	18.8

Table 6: Results on MQR TEST set. The "Ill-formed" shows metric scores for the questions in TEST without rewriting. The next portion shows results for models trained on the TRAIN portion of MQR. The lower portion shows results for methods using other models and/or datasets.

Training Dataset	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
MQR TRAIN	22.1	59.8	32.2	56.6	36.4
MQR TRAIN + ⟨well-formed, well-formed⟩ pairs	21.1	**61.4**	32.1	**58.0**	**36.8**
MQR TRAIN + Quora	**23.6**	60.5	**33.4**	57.5	**36.8**
MQR TRAIN + Paralex	21.7	58.3	31.3	55.3	35.7
MQR TRAIN + Quora + Paralex	23.1	60.3	33.0	57.2	36.7

Table 7: Results showing how additional training data affects performance for the transformer model.

In following sections, when a transformer model is used, we follow the same setting as described above.

LSTM Sequence to Sequence Model with Attention. We use the attention mechanism proposed by (Luong, Pham, and Manning 2015). We use the Tensor2Tensor implementation (Vaswani et al. 2018) with their provided Luong Attention hyperparameter settings. We set batch size to 4096. The hidden size is 1000 and we use 4 LSTM hidden layers following (Luong, Pham, and Manning 2015).

Methods Built from Other Resources

We also benchmark other methods involving different training datasets and models. All the methods in this subsection use transformer models.

Round Trip Neural Machine Translation. Round trip neural machine translation is an effective approach for question or sentence paraphrasing (Mallinson, Sennrich, and Lapata 2017; Dong et al. 2017; Iyyer et al. 2018). It first translates a sentence to another pivot language, then translates it back to the original language. We consider the use of both German (De) and French (Fr) as the pivot language, so we require translation systems for En↔De and En↔Fr.

The English-German translation models are trained on WMT datasets, including News Commentary 13, Europarl v7, and Common Crawl, and evaluated on newstest2013 for early stopping. On the newstest2013 dev set, the En→De model reaches a BLEU-4 score of 19.6, and the De→En model reaches a BLEU-4 score of 24.6.

The English-French models are trained on Common Crawl 13, Europarl v7, News Commentary v9, Giga release 2, and UN doc 2000. On the newstest2013 dev set, the En→Fr model reaches a BLEU-4 score of 25.6, and the Fr→En model reaches a BLEU-4 score of 26.1.

Grammatical Error Correction (GEC). As some ill-formed questions are not grammatical, we benchmark a state-of-the-art grammatical error correction system on this task. We use the system of (Lichtarge et al. 2019), a GEC ensemble model trained from Wikipedia edit histories and round trip translations.

Paraphrase Generator Trained on ParaNMT. We also train a paraphrase generation model on a subset of the ParaNMT dataset (Wieting and Gimpel 2018), which was created automatically by using neural machine translation to translate the Czech side of a large Czech-English parallel corpus. We use the filtered subset of 5M pairs provided by the authors. For each pair of paraphrases (S1 and S2) in the dataset, we train the model to rewrite from S1 to S2 and also rewrite from S2 to S1. We use the MQR DEV set for early stopping during training.

Results

Table 6 shows the performance of the models and methods described above. Among these methods models trained on MQR work best. GEC corrects grammatical errors and spelling errors, so it also improves the question quality in

	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Transformer (MQR + Quora)	23.6	60.5	33.4	57.5	36.8
GEC	13.1	52.4	24.4	47.5	34.4
GEC → Transformer (MQR + Quora)	24.8	60.2	33.9	57.3	36.8
Transformer (MQR + Quora) → GEC	**26.3**	**61.0**	**35.4**	**58.1**	**37.3**
Transformer (MQR) → Transformer (MQR)	20.4	55.8	29.2	52.5	35.1

Table 8: Methods combining transformer trained on MQR + Quora with GEC. "A → B" means running method A followed by method B on method A's output.

	Spelling	Grammar	Explicit	Semantics wrt. ill-formed	Semantics wrt. well-formed
Ill formed	0.31	0.41	0.61	-	-
GEC	0.39	0.56	0.59	1.00	0.84
Transformer (MQR + Quora)	0.96	0.75	1.00	0.67	0.56
Transformer (MQR + Quora) → GEC	0.96	0.91	1.00	0.71	0.63

Table 9: Results of human evaluation of three models on 75 test examples.

rewriting. Round trip neural machine translation is a faithful rewrite of the questions, and it naturally corrects some spelling and grammatical errors during both rounds of translation due to the strong language models present in the NMT models. However, it fails in converting commands and statements into questions.

The paraphrase generator trained on ParaNMT does not perform well, likely because of domain difference (there are not many questions in ParaNMT). It also is unlikely to convert non-question sentences into explicit questions.

Additional Training Data

We consider two additional data resources to improve question rewriting models.

The first resource is the Quora Question Pairs dataset.[7] This dataset contains question pairs from Quora, an online question answering community. Some question pairs are marked as duplicate by human annotators and other are not. We consider all Quora Question Pairs (Q1 and Q2) marked as duplicate as additional training data. We train the model to rewrite from Q1 to Q2 and also from Q2 to Q1. This gives us 298,364 more question pairs for training.

The second resource is the Paralex dataset (Fader, Zettlemoyer, and Etzioni 2013). The questions in Paralex are scraped from WikiAnswers,[8] where questions with similar content are clustered. As questions in the Paralex dataset may be noisy, we use the annotation from (Faruqui and Das 2018). Following their standard, we treat all questions with scores higher than 0.8 as well-formed questions. For each well-formed question, we take all questions in the same Paralex question cluster and construct pairs to rewrite from other questions in the cluster to the single well-formed question. This gives us 169,682 extra question pairs for training.

We also tried adding "identity" training examples in which the well-formed questions from the MQR TRAIN set are repeated to form a question pair.

The results of adding training data are summarized in Table 7. Adding the identity pairs improves the ROUGE and METEOR scores, which are focused more on recall, while harming BLEU, which is focused on precision. We hypothesize that adding auto-encoding data improves semantic preservation, which is expected to help the recall-oriented metrics. Adding Quora Question Pairs improves performance on TEST but adding Paralex pairs does not. The reason may stem from domain differences: WikiAnswers (used in Paralex) is focused on factoid questions answered by encyclopedic knowledge while Quora and Stack Exchange questions are mainly answered by community contributors. Semantic drift occurs more often in Paralex question pairs as Paralex is constructed from question clusters, and a cluster often contains more than 5 questions with significant variation.

Combining Methods

In addition to the aforementioned methods, we also try combining multiple approaches. Table 8 shows results when combining GEC and the Quora-augmented transformer model. We find that combining GEC and a transformer question rewriting model achieves better results than each alone. In particular, it is best to first rewrite the question using the transformer trained on MQR + Quora, then run GEC on the output.

We also tried applying the transformer (trained on MQR) twice, but it hurts the performance compared to applying it only once (see Table 8).

Human Evaluation

To better evaluate model performance, we conduct a human evaluation on the model rewritten questions following the same guidelines from the "Dataset Quality" subsection. Among the 300 questions annotated earlier, we chose the ill-formed questions from the TEST split, which yields 75 questions. We evaluate questions rewritten by three methods (Transformer (MQR + Quora), GEC, and Transformer

[7] https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs

[8] http://wiki.answers.com/

model	question	S	G	E	Semantics
Ill-formed	best way of widening butcherblock countertop?	0	0	0	1
Well-formed	What's the best way to widen a butcherblock countertop?	1	1	1	-
Trans.	How can I widen a butcherblock countertop?	1	1	1	0
LSTM	What is the best way of widening butcherblock countertop?	1	0	1	1
GEC	best way of widening butcherblock countertop?	0	1	0	1
Round trip NMT (Pivot: De)	best way to extend the racquet counter pole?	0	1	0	0
Round trip NMT (Pivot: Fr)	What is the best way to expand the butcherblock?	1	1	1	0
ParaNMT	the best way to expand the countertop of the butcher ?	1	1	0	1
Trans. (MQR + Quora)	What is the best way of widebutcherblock countertop?	0	0	0	0
Trans. (MQR + Quora) → GEC	What is the best way of widebitcherblock countertop?	0	0	0	0
Ill-formed	drawing polygons from python console	0	1	0	1
Well-formed	How to draw polygons from the python console?	1	1	1	-
Trans.	How to draw polygons from a python console?	1	1	1	1
LSTM	How can I draw polygons from a Python console?	1	1	1	1
GEC	drawing polygons from python console	0	0	0	1
Round trip NMT (Pivot: De)	Drawing polygons from the Python console	0	1	0	1
Round trip NMT (Pivot: Fr)	polygons of the python console	0	1	0	0
ParaNMT	drawing polygons from python console	0	0	0	1
Trans. (MQR + Quora)	How to draw polygons from python console?	1	0	1	1
Trans. (MQR + Quora) → GEC	How to draw polygons from a python console?	1	1	1	1

Table 10: Examples of ill-formed question rewritten by models with human annotations. (S = Spelling, G = Grammar, and E = Explicit) The last column shows semantic equivalence with the well-formed questions.

(MQR + Quora) → GEC), and ask annotators to determine the qualities of the rewritten questions. To understand if question meanings change after rewriting, we also annotate whether a model rewritten question is semantically equivalent to the ill-formed question or equivalent to the well-formed one.

Table 9 shows the annotations from two annotators. When the two annotators disagree, a judge makes a final decision. Note that the examples annotated here are a subset of those annotated in Table 4, so the first row is different from the ill-formed questions in Table 4. According to the annotations, the GEC method slightly improves the question quality scores. Although Table 6 shows that GEC improves the question quality by some automatic metrics, it simply corrects a few grammatical errors and the rewritten questions still do not meet the standards of human annotators. However, the GEC model is good at preserving question semantics.

The Transformer (MQR + Quora) model and Transformer (MQR + Quora) → GEC excel at improving question quality in all three aspects, but they suffer from semantic drift. This suggests that future work should focus on solving the problem of semantic drift when building question rewriting models.

Table 10 shows two example questions rewritten by different methods. The questions rewritten by GEC remain unchanged but are still of low quality, whereas ParaNMT and round trip NMT make a variety of changes, resulting in large variations in question quality and semantics. Methods trained on MQR excel at converting ill-formed questions into explicit ones (e.g., adding "What is" in the first example and "How to" in the second example), but sometimes make grammatical errors (e.g., Trans. (MQR + Quora) misses "a" in the second example). According to Table 8, combining neural models trained on MQR and GEC achieves the best results in automatic metrics. However, they still suffer from semantic drift. In the first example of Table 10, the last two rewrites show significant semantic mistakes, generating nonexistent words "widebutcherblock" and "widebitcherblock".

Conclusion and Future Work

We proposed the task of question rewriting and produced a novel dataset MQR to target it. Our evaluation shows consistent gains in metric scores when using our dataset compared to systems derived from previous resources. A key challenge for future work is to design better models to rewrite ill-formed questions without changing their semantics. Alternatively, we could attempt to model the process whereby question content changes. Sometimes community members do change the content of questions in online forums. Such rewrites typically require extra context information, such as the question description. Additional work will be needed to address this context-sensitive question rewriting task.

Acknowledgments

We thank Shankar Kumar, Zi Yang, Yiran Zhang, Rahul Gupta, Dekang Lin, Yuchen Lin, Guan-lin Chao, Llion Jones, and Amarnag Subramanya for their helpful discussions and suggestions.

References

Androutsopoulos, I., and Malakasiotis, P. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research* 38:135–187.

Banerjee, S., and Lavie, A. 2005. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*.

Bannard, C., and Callison-Burch, C. 2005. Paraphrasing with bilingual parallel corpora. In *Proc. of ACL*.

Barzilay, R., and Lee, L. 2003. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proc. of NAACL*.

Buck, C.; Bulian, J.; Ciaramita, M.; Gajewski, W.; Gesmundo, A.; Houlsby, N.; and Wang, W. 2018. Ask the right questions: Active question reformulation with reinforcement learning. *ICLR*.

Chen, M.; Tang, Q.; Wiseman, S.; and Gimpel, K. 2019. Controllable paraphrase generation with a syntactic exemplar. In *Proc. of ACL*.

Cho, E.; Xie, H.; and Campbell, W. M. 2019. Paraphrase generation for semi-supervised learning in NLU. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*.

De Bona, F.; Riezler, S.; Hall, K.; Ciaramita, M.; Herdağdelen, A.; and Holmqvist, M. 2010. Learning dense models of query similarity from user click logs. In *Proc. of NAACL*.

Dong, L.; Mallinson, J.; Reddy, S.; and Lapata, M. 2017. Learning to paraphrase for question answering. In *Proc. of EMNLP*.

Eisenstein, J. 2013. What to do about bad language on the internet. In *Proc. of NAACL*.

Fader, A.; Zettlemoyer, L.; and Etzioni, O. 2013. Paraphrase-driven learning for open question answering. In *Proc. of ACL*.

Faruqui, M., and Das, D. 2018. Identifying well-formed natural language questions. In *Proc. of EMNLP*.

Gupta, A.; Agarwal, A.; Singh, P.; and Rai, P. 2018. A deep generative framework for paraphrase generation. In *Proc. of AAAI*.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Comput.* 9(8):1735–1780.

Iyyer, M.; Wieting, J.; Gimpel, K.; and Zettlemoyer, L. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proc. of NAACL*, 1875–1885.

Jones, R.; Rey, B.; Madani, O.; and Greiner, W. 2006. Generating query substitutions. In *Proc. of WWW*, 387–396. ACM.

Kumar, A.; Dandapat, S.; and Chordia, S. 2018. Translating web search queries into natural language questions. In *Proc. of LREC*.

Li, Z.; Jiang, X.; Shang, L.; and Li, H. 2018. Paraphrase generation with deep reinforcement learning. In *Proc. of EMNLP*, 3865–3878.

Li, Z.; Jiang, X.; Shang, L.; and Liu, Q. 2019. Decomposable neural paraphrase generation. *arXiv preprint arXiv:1906.09741*.

Lichtarge, J.; Alberti, C.; Kumar, S.; Shazeer, N.; Parmar, N.; and Tong, S. 2019. Corpora generation for grammatical error correction. In *Proc. of NAACL*.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Luong, T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. In *Proc. of EMNLP*.

Madnani, N., and Dorr, B. J. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics* 36(3).

Malakasiotis, P., and Androutsopoulos, I. 2011. A generate and rank approach to sentence paraphrasing. In *Proc. of EMNLP*.

Mallinson, J.; Sennrich, R.; and Lapata, M. 2017. Paraphrasing revisited with neural machine translation. In *Proc. of EACL*.

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

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proc. of ACL*.

Sproat, R.; Black, A. W.; Chen, S.; Kumar, S.; Ostendorf, M.; and Richards, C. 2001. Normalization of non-standard words. *Computer speech & language* 15(3):287–333.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proc. of NIPS*.

Vaswani, A.; Bengio, S.; Brevdo, E.; Chollet, F.; Gomez, A. N.; Gouws, S.; Jones, L.; Kaiser, L.; Kalchbrenner, N.; Parmar, N.; Sepassi, R.; Shazeer, N.; and Uszkoreit, J. 2018. Tensor2tensor for neural machine translation. *CoRR* abs/1803.07416.

Wieting, J., and Gimpel, K. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proc. of ACL*.

Zhang, W. V., and Jones, R. 2007. Comparing click logs and editorial labels for training query rewriting. In *WWW 2007 Workshop on Query Log Analysis: Social And Technological Challenges*.

Zhang, W. V.; He, X.; Rey, B.; and Jones, R. 2007. Query rewriting using active learning for sponsored search. In *Proc. of SIGIR*.