- Answers can contain up to a maximum of 30 words.
- When the answer is not a text, provide the source URL only (e.g., a geo-location on a map or a music video link).

C Pitfalls of the Query Rewriting Metrics

Our evaluation results show that the text similarity metrics, such as ROUGE and USE, often fall short to reflect semantic similarity in case of lexical paraphrases. Retrieval-based metrics, such as Recall@10, are able to demonstrate better correlation with human judgement. However, retrieval-based metrics are more expensive to compute since it requires an API call for every query. Also, they rely on the underlying collection as well as the ability of the search engine to handle paraphrases. Our experiments show that text similarity metrics, however flawed, are still able to provide a good proxy for quickly assessing QR performance and are suitable for comparing models in the development phase during parameter tuning. Retrieval-based metrics are useful to better approximate human judgement but can be computed for the best models only that were pre-selected using text similarity metrics.

ROUGE-1 R metrics provides a very rough estimate of the model performance by counting the number of words missing from the generated question rewrite in comparison with the ground truth rewrite and does not have any mechanism to distinguish which words are more crucial than others. As a result, a question missing only a single letter will receive the same score as a question missing one of its most informative words. For example, ROUGE("When is Robert Downey *Jr* birthday", "When is Robert Downey *Jrs* birthday") = ROUGE("When did Gabriel Garcia die", "When did Gabriel Garcia *Marquez* die") = 0.75.

USE is more sensitive to such variations and can better pick up on the character-level similarities: compare to USE("When is Robert Downey *Jr* birthday", "When is Robert Downey *Jrs* birthday")=0.96 and USE("When did Gabriel Garcia die", "When did Gabriel Garcia *Marquez* die") = 0.91.

Web search results, while most accurately correlates with human judgment, also reflect sensitivity of the retrieval algorithm to the query formulation as well as the collection-specific selectivity of the query terms. The resulting scores for our sample rewrites are R@10("When is Robert Downey *Jr* birthday", "When is Robert Downey *Jrs* birthday")=0.6 and R@10("When did Gabriel Garcia die", "when did Gabriel Garcia *Marquez* die") = 0.78.

D Examples of Query Rewrites

In Table 9 we show sample question rewrites from top 3 QR models along with conversational context.

E Examples of Answers Found

In Table 10 we provide two sample answers found by the baseline model. In the first example, the baseline system picked the same passage as the human annotator, but extracted a different answer span from this passage. In the second example, the baseline system picked a different passage than the human annotator. While the produced answers are not exact matches to the corresponding human answers, we consider them as correct upon manual inspection. This observation that a single question in QReCC can have multiple correct answers poses an important challenge for standard approaches to QA evaluation.

Table 9: A sample of conversation snippets from the QReCC test set with the question rewrites produced by Transformer++, CopyTransformer, and GECOR models.

Conversational context	**Transformer++**	**CopyTransformer**	**GECOR**
Q1: Did Nadia Comăneci win any Gold medals in the Olympics? *A1:* Nadia Comăneci is a five-time Olympic gold medalist. *Q2:* What about Silver?	Did Nadia Comăneci win any Silver medals?	Did Nadia win Silver medals?	What about Silver medals?
Q1: What is range in statistics? *A1:* The Range is the difference between the lowest and highest values. *Q2:* Describe some different metrics to interpret it.	Describe some different metrics to interpret range in statistics.	Describe some different metrics to interpret range in statistics.	Describe some different metrics to interpret range.
Q1: What is nominal GDP? *A1:* Nominal GDP is GDP evaluated at current market prices and includes all of the changes in market prices that have occurred during the current year due to inflation or deflation. *Q2:* What is the difference with real?	What is the difference between nominal and real GDP?	What is the difference with real GDP?	What is the difference with real GDP?
Q1: Tell me about lavender plants? *A1:* Lavandula is a genus of 47 known species of flowering plants in the mint family, Lamiaceae. It is native to the Old World and is found from Cape Verde and the Canary Islands, Europe across to northern and eastern Africa, the Mediterranean, southwest Asia to southeast India. *Q2:* What are the different types?	What are the different types of lavender plants?	What are the different types of plants?	What are the different types of plants?

Table 10: A sample of answers produced by our end-to-end baseline for conversational QA. The baseline model can also produce relevant answers using spans that differ from the answers provided by the human annotators.

Human re-written question	What are the educational requirements required to become a physician's assistant?
URL	https://www.geteducated.com/careers/how-to-become-a-physician-assistant
Predicted URL	https://www.geteducated.com/careers/how-to-become-a-physician-assistant
Human passage	. . . In most cases, a physician assistant will need a master's degree from an accredited institution (two years of post-graduate education after completing a four-year degree). . . . Most applicants to PA education programs will not only have four years of education, they will also have at least a year of medical experience. . . . five steps to becoming a PA: Complete your bachelor's degree (a science or healthcare related major is usually best); Gain experience either working or volunteering in a healthcare setting; Apply to ARC-PA accredited programs; Complete a 2-3 year, master's level program; Pass the PANCE licensing exam.
Found passage	(Same as human passage.)
Human answer	Complete your bachelor's degree (a science or healthcare related major is usually best); Gain experience either working or volunteering in a healthcare setting; Apply to ARC-PA accredited physician assistant programs; Complete a 2-3 year, master's level PA program;
Baseline model answer	a physician assistant will need a master's degree from an accredited institution (two years of post-graduate education after completing a four-year
Answer F1	15.38
Human re-written question	What tools were used in the neolithic event?
URL	https://sciencing.com/list-neolithic-stone-tools-8252604.html
Predicted URL	https://stmuhistorymedia.org/neolithic-era-technology-advances-and-beginnings-of-agriculture
Human passage	. . . By the time the Neolithic came around, hand axes had fallen out of favor . . . scientists consider the creation of all these tools a sign of early human ingenuity. Scrapers Scrapers are one of the original stone tools, found everywhere where people settled, . . . Blades While a scraper can be used for cutting into an animal, a longer, thinner blade can be inserted deeper into a carcass, . . . Arrows and Spearheads Arrows and spearheads are a more sophisticated shape than simple scrapers and blades. . . . Axes The polished stone ax is considered one of the most important developments of the Neolithic era. . . . Adzes The adze is a woodworking tool. . . . Hammers and Chisels Chisels were made by attaching a sharp piece of stone to the end of a sturdy stick . . .
Found passage	. . . The Neolithic Age was a period in the development of human technology, beginning about 10,000 BCE, in some parts of the Middle East, and later in other parts of the world, and ending between 4,500 and 2,000 BCE. . . . Hunting also became much easier to accomplish with the introduction new of stone tools. The most common tools used were daggers and spear points, used for hunting, and hand axes, used for cutting up different meats, and scrappers, which were used to clean animal hides.
Human answer	Scrapers. Scrapers are one of the original stone tools, found everywhere where people settled, long before the Neolithic Age began. ...Blades. ...Arrows and Spearheads. ...Axes. ...Adzes. ...Hammers and Chisels.
Baseline model answer	The most common tools used were daggers and spear points, used for hunting, and hand axes
Answer F1	19.05