QED: A Framework and Dataset for Explanations in Question Answering

Matthew Lamm\textsuperscript{1}, Jennimaria Palomaki\textsuperscript{2}, Chris Alberti\textsuperscript{2}, Daniel Andor\textsuperscript{2}, Eunsol Choi\textsuperscript{3}, Livio Baldini Soares\textsuperscript{2}, Michael Collins\textsuperscript{2}

\textsuperscript{1}Department of Linguistics, Stanford University, United States
\textsuperscript{2}Google Research, United States
\textsuperscript{3}Department of Computer Science, The University of Texas at Austin, United States
\{mrlamm, jpalomaki, chrisalberti, andor, liviobs, mjcollins\}@google.com, eunsol@cs.utexas.edu

Abstract

A question answering system that in addition to providing an answer provides an explanation of the reasoning that leads to that answer has potential advantages in terms of debuggability, extensibility, and trust. To this end, we propose QED, a linguistically informed, extensible framework for explanations in question answering. A QED explanation specifies the relationship between a question and answer according to formal semantic notions such as referential equality, sentencehood, and entailment. We describe and publicly release an expert-annotated dataset of QED explanations built upon a subset of the Google Natural Questions dataset, and report baseline models on two tasks—post-hoc explanation generation given an answer, and joint question answering and explanation generation. In the joint setting, a promising result suggests that training on a relatively small amount of QED data can improve question answering. In addition to describing the formal, language-theoretic motivations for the QED approach, we describe a large user study showing that the presence of QED explanations significantly improves the ability of untrained raters to spot errors made by a strong neural QA baseline.

1 Introduction

Question answering (QA) systems can enable efficient access to the vast amount of information that exists as text (Rajpurkar et al., 2016; Kwiatkowski et al., 2019; Clark et al., 2019; Reddy et al., among others). Modern neural systems have made tremendous progress in QA accuracy in recent years (Devlin et al., 2019). However, they generally give no explanation or justification of how they arrive at an answer to a question. Models that in addition to providing an answer can explain their reasoning may have significant benefits pertaining to trust and debuggability (Doshi-Velez and Kim, 2017; Ehsan et al., 2019).

Critical questions then, are what constitutes an explanation in question answering, and how we can enable models to provide such explanations. In an effort to make progress on these questions, in this paper we (1) introduce QED, a linguistically grounded definition of explanations for extractive QA; and (2) describe an expert-annotated corpus of QED annotations based on the Natural Questions (Kwiatkowski et al., 2019) dataset. The QED corpus has been released publicly.

Figure 1 shows a QED example. Given a question and a passage, QED represents an explanation as a combination of discrete, human-interpretable steps: (1) identification of a sentence implying an answer to the question, (2) identification of noun phrases in both the question and answering sentence that refer to the same thing, and (3) confirmation that the predicate in the sentence entails the predicate in the question once referential equalities are abstracted away.

This choice of explanation makes use of core semantic relations—referential equality and entailment—and thus has well-understood formal properties. In addition, we found that this way of decomposing explanations has high coverage

\textsuperscript{1}QED stands for the Latin “quod erat demonstrandum” or “that which was to be shown”.
\textsuperscript{2}https://github.com/google-research-datasets/QED.
Figure 1: QED explanations decompose the question-passage relationship in terms of referential equality and predicate entailment.

(77% on the Natural Questions corpus) and can be readily extended to other forms of question answering. (See Section 6.) Since QED decomposes the QA process into distinct subproblems, we also believe that it should enable research directions aimed at extending or improving upon extant QA systems.

In what follows, we present a definition of QED explanations. We then describe the dataset of QED annotations (7638/1353 train/dev examples), including discussion of the distribution of linguistic phenomena exhibited in the data. We move to propose four potential tasks, of varying complexity, related to the QED framework, and use the QED annotations to train and evaluate different models on two of these. Additionally, we describe a rater study which shows how the presence of QED explanations can help users identify errors made by an automated QA system.

2 Annotation Definition

We now describe the form of QED annotations. The treatment in this section is somewhat informal; for formal definitions see Appendix A.

2.1 Basic Definitions

We will use the following example to illustrate the approach:

<table>
<thead>
<tr>
<th>Question: who wrote the film Howl’s Moving Castle?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passage: Howl’s Moving Castle is a 2004 Japanese animated fantasy film written and directed by Hayao Miyazaki. It is based on the novel of the same name, which was written by Diana Wynne Jones. The film was produced by Studio Ghibli. Answer: Hayao Miyazaki</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(1) Single Sentence Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Howl’s Moving Castle is a 2004 Japanese animated fantasy film written and directed by Hayao Miyazaki.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(2) Referential Equality</th>
</tr>
</thead>
<tbody>
<tr>
<td>the film Howl’s moving castle = Howl’s Moving Castle</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(3) Entailment</th>
</tr>
</thead>
<tbody>
<tr>
<td>X is a 2004 Japanese animated fantasy film written and directed by ANSWER. ANSWER wrote X.</td>
</tr>
</tbody>
</table>

The annotator is presented with a question/passage pair. Annotation then proceeds in the following four steps:

(1) **Single Sentence Selection.** The annotator identifies a single sentence in the passage that entails an answer to the question assuming that coreference and bridging anaphora (see later in this section) have been resolved in the sentence.

In the above example, the following sentence entails an answer to the question, and would be selected by the annotator:

```
Its official capacity is 107,601.
```

(2) **Answer Selection.** The annotator highlights a short answer span (or spans) in the answer sentence. In the above example the annotator would mark the following (answer shown with [ = A ... ]): (3) **Identification of Question–Sentence Noun Phrase Equalities.** The annotator marks referentially equivalent noun phrases, or noun phrases that refer to the same thing, in the question and the answer sentence. This includes reference not only to individuals and other proper nouns, but also to generic concepts.

In our example the annotator would mark the following two noun phrases (marked with the [ = I ... ] annotations) as referentially equivalent:

```
Its official capacity is [ = A 107,601 ].
```

(4) **If it is not possible to find a sentence that satisfies these properties—typically because the answer requires inference beyond coreference/bridging that involves multiple sentences—the annotator marks the example as not possible. See Section 3.
how many seats in [=1 university of michigan stadium] 
[=1 Its] official capacity is [=A 107,601].

(4) Extraction of an Entailment Pattern. As a final, automatic step, an entailment pattern can be extracted from the annotated example by abstracting over referentially equivalent noun phrases. In the above example the entailment pattern would be as follows:

how many seats in X
X’s official capacity is [=A 107,601].

2.2 Two Extensions

There are two extensions to the above approach, coreference in answers, and bridging in referential equalities:

Coreference in Answers. Consider the following example:

<table>
<thead>
<tr>
<th>Question: who won wimbledon in 2019</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passage: Simona Halep is a female tennis player. She won Wimbledon in 2019.</td>
</tr>
</tbody>
</table>

In this case the single sentence She won Wimbledon in 2019 would be selected by the annotator in step 1, as once coreference is resolved, this entails the answer to the question. The QED annotation would be as follows:

who won [=1 wimbledon] in [=2 2019]

[=A She (=C Simona Halep)] won [=1 Wimbledon] in [=2 2019]

In this case the answer “She”—the substring in the original sentence—is not sufficient, as it involves an unresolved anaphor. Because of this, the annotator would mark the fact that “She” refers to “Simona Halep” earlier in the passage, using the (= C . . . ) notation.

Bridging in Referential Equalities. Bridging anaphora (Clark, 1975) are frequently encountered in the QA passages in our data, and in Wikipedia more broadly. Consider the following:

<table>
<thead>
<tr>
<th>Question: who won america’s got talent season 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passage: The 11th season of America’s Got Talent, an American talent show competition, began broadcasting in the United States during 2016. Grace VanderWaal was announced as the winner on September 14, 2016.</td>
</tr>
</tbody>
</table>

It is clear from the context surrounding the sentence “Grace VanderWaal was announced as the winner on September 14, 2016” that the noun phrase “the winner” refers to “the winner of America’s Got Talent Season 11”, and hence the sentence provides an answer to the question. It is helpful to imagine that there is an implicit prepositional phrase “of America’s Got Talent Season 11” modifying “the winner”. In this case the annotation would be the following:

who won [=1 america’s got talent season 11]

[=A Grace VanderWaal] was announced as [=B the winner (of =1)] on September 14, 2016.

Here the annotation [=B the winner (of =1)] indicates that the phrase marked [=1 . . . ] in the query is a bridged modifier to the phrase the winner, through the preposition “of”.

Sometimes, there is no phrase like “the winner” above, but the referent is clearly an implicit argument of the supporting sentence. In this case we treat it as a bridge into the entire sentence.

2.3 A Note on Terminology

In defining QED we use the terms “predicate” and “entailment” in ways that may seem unfamiliar, but are not unrelated to the typical senses of those terms in linguistics. Canonically speaking, one thinks of a predicate as the semantic correlate of a verb in a sentence, and usually containing information about its argument structure. By taking a less structured notion of the term, as everything in a sentence surrounding a set of salient referring expressions, we are able to strike a balance between completely unstructured text, and more elaborate, structured representations that tend to be brittle.

The sense of entailment we intend then follows from this definition. A sentence entails an answer to a question if, having resolved and abstracted away referential equalities between the two, one can identify an answer to the question in the sentence.

3 QED Annotations for the Natural Questions

We now describe QED annotations over the Natural Questions (NQ) dataset (Kwiatkowski et al., 2019). We first describe the annotation process,
then describe agreement statistics and statistics of types of referential expressions. For discussion of the assumptions we make and future extensions to QED, please see Section 6.3.

We focus on questions in the NQ corpus that have both a passage and short answer marked by the NQ annotator. We exclude examples where the passage is a table. A QED annotator was presented with a question/paragraph pair. Before performing the core QED annotation, annotators first determine whether: (1) there is a valid short answer within the paragraph (note that they can overrule the original NQ judgment), and there is a valid QED explanation for that answer; (2) there is a valid short answer within the paragraph, but there is no valid QED explanation for that answer; or (3) there is no valid short answer within the passage (hence the original NQ annotation is judged to be an error). Ten percent of all examples fell into category (3). Of the remaining 90% of examples that contained a correct short answer, 77% fell into category (1), and 23% fell into category (2).

Three QED annotators5 annotated 7638 training examples (5154/1702/782 in categories 1/2/3 respectively), and 1353 dev examples (1019/183/151 in categories 1/2/3), without replication. We estimate that annotators averaged approximately 2 minutes per instance. Additionally, early stages of annotation consisted of regular adjudication among annotators to establish a consensus on QED’s guidelines.

3.1 Agreement Statistics

All three annotators marked a common set of 100 examples drawn from the development set. We compute average pairwise agreement by comparing each annotator against the other two, and averaging across all pairs. Average classification of instances was 73.9%. If this seems low, it is because one annotator was more conservative interpreting QED’s single sentence assumption, and pairwise accuracy breakdown was thus 81.2/72.3/68.1%. Given the high number of “debatable” instances reported in the Natural Questions paper, this divergence is unsurprising, however. Average pairwise F1 on mention identification/mention alignment, conditioned on both annotators labeling instances as amenable to QED, was 88.4 and 84.1, respectively.

3.2 Types of Referential Expressions

The referential equality annotations are a major component of QED. Figure 2 shows some full QED examples from the corpus, and Figure 3 shows some example referential equalities from the corpus. In this section, in an effort to gain insight about the types of phenomena present, we describe statistics on types of referential equalities. We subcategorize referring expressions into the following types.6

Proper Names Examples are “How I met your Mother” or “the cbs television sitcom how i met your mother”.

Non-Anaphoric Definite NPs These are expressions such as “the president of the United States” or “the next Maze Runner film”. The majority involve one or more common nouns (e.g., “president”, “film”) together with a proper name, thereby defining a new entity that is in some sense a “derivative” of the underlying proper name.

Anaphoric Definite NPs These are definite NPs, most often from within the passage rather than the question, that require context to be interpreted. Examples are “the series” referring to an earlier mention of “the Vampire Diaries” within the passage, or “the winner” referring to “the winner of America’s got Talent Season 11”.

Generics Examples are “a dead zone” in the question “what causes a dead zone in the ocean”, or “Dead zones” in the passage sentence “Dead zones are low-oxygen areas caused by . . . ”.

Pronouns Examples are it, they, he, she.

Bridging Referential expressions in the passage sentence that use bridging (see Section 2.2).

Miscellaneous All referential expressions not included in the categories above.

Table 1 shows the frequency distribution of per-instance referential equality counts. Figure 4 shows an analysis of 100 referential equality annotations from QED, with a breakdown by type of referring expression in the question and passage. Proper names, non-anaphoric definites,

5Three of the authors of this paper.

6For formal discussion, see Carlson (1977), Krifka (2003), Abbott (2004), and Mikkelsen (2011), among others.
and generics dominate expression types in the question (73, 16, and 6 examples, respectively). Expressions in the sentence are more diverse, with a much greater proportion of anaphoric definites, pronouns, and bridging examples (21, 9, and 5 cases, respectively). Finally, as an indication of the difficulty of the referential equality task, we note that in only 12% of all referential equalities in the 100 examples in Figure 4 is there an exact string match (after lower-casing of both question and passage) between the question and passage referential expression.

### Figure 2: Examples from the QED dataset, grouped according to different types of referential equalities.

<table>
<thead>
<tr>
<th>Type</th>
<th>Question</th>
<th>Wikipedia page</th>
<th>Passage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pronominal reference</td>
<td>Question: how many blocks in the great pyramid of giza?</td>
<td>Great_Pyramid_of_Giza</td>
<td>Based on these estimates, building the pyramid in 20 years would involve installing approximately 800 tonnes of stone every day. Additionally, since it consists of an estimated 2.3 million blocks, completing the building in 20 years would involve moving an average of more than 12 of the blocks into place each hour, day and night. The first precision measurements of the pyramid were made by Egyptologist Sir Flinders Petrie in 1880–82 and published as The Pyramids and Temples of Gizeh. Almost all reports are based on his measurements.</td>
</tr>
<tr>
<td>Inexact match</td>
<td>Question: where does the term sixes and sevens originate</td>
<td>At_sixes_and_sevens</td>
<td>An ancient dispute between the Merchant Taylors and Skinners livery companies, is probably the origin of the phrase. The two trade associations, both founded in the same year (1327), argued over sixth place in the order of precedence. In 1484, after more than a century and a half of bickering, the Lord Mayor of London Sir Robert Billesden ruled that at the feast of Corpus Christi, the companies would swap between sixth and seventh place and feast in each other's halls.</td>
</tr>
<tr>
<td>Answer bridging/coreference</td>
<td>Question: what is Whitney Houston's name</td>
<td>Cissy_Houston</td>
<td>Emily “Cissy” Houston, (née Drinkard; born September 30, 1933) is an American soul and gospel singer. After a successful career singing backup for such artists as Dionne Warwick, Elvis Presley and Aretha Franklin, Houston embarked on a solo career, winning two Grammy Awards for her work. Houston is the mother of singer Whitney Houston, grandmother of Whitney's daughter, Bobbi Kristina Brown, aunt of singers Dionne and Dee Dee Warwick, and a cousin of opera singer Leontyne Price.</td>
</tr>
<tr>
<td>Entity bridging</td>
<td>Question: who sang the national anthem at the first game of 2017 world series</td>
<td>2017_World_Series</td>
<td>Game 1: The ceremonial first pitch was thrown out by members of former Dodger Jackie Robinson's family, including his widow Rachel. The game marked the 45th anniversary of Robinson's death, and the 2017 season was the 70th anniversary of his breaking of the baseball color line. (at ...), Keith Williams Jr., a gospel singer, performed &quot;The Star-Spangled Banner&quot;, the national anthem.</td>
</tr>
<tr>
<td>Generic reference</td>
<td>Question: what is the function of a paints binder</td>
<td>Paint</td>
<td>The binder is the film-forming component of paint. It is the only component that is always present among all the various types of formulations. Many binders are too thick to be applied and must be thinned. The type of thinner, if present, varies with the binder.</td>
</tr>
</tbody>
</table>

4 Tasks and Models

The QED data, which we release publicly, can be used as part of wide range of QA tasks and models. After discussing some of these tasks, we assess how well two recent neural architectures, one structured and one sequence-to-sequence, perform on two of them.

4.1 QED-based Tasks

Each QED example is a \((q, d, c, a, e)\) tuple where \(q\) is a question from the NQ dataset, \(d\) is a Wikipedia
**Table 1: Referential link count frequency distribution in a random sample of 1000 instances.** When there are 0 links, the explanation consists only of a selected sentence.

<table>
<thead>
<tr>
<th>Referential Link Count</th>
<th>Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>54</td>
</tr>
<tr>
<td>1</td>
<td>649</td>
</tr>
<tr>
<td>2</td>
<td>294</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
</tbody>
</table>

**Figure 4:** Counts for 100 randomly drawn referential equality annotations from the QED corpus, sub-categorized by expression type in the question (Qu.) and passage (Ps.). P/N/A/G/Pn/B/M refer to Proper/Def (non-ana)/Def(ana)/Generic/Pronoun/Bridge/Misc.

**Task 2: Joint Answer and Explanation Prediction** Given a \((q, d, c)\) triple, predict \((\hat{a}, \hat{e}) = f(q, d, c)\), where \(f\) is a function that maps it to a short-answer/explanation pair. We might, for example, define \(f(q, d, c) = \arg\max_{a, e} p(a, e | q, d, c; \theta)\) under some model \(p(\ldots)\). The evaluation measure is \(\sum_{(q,d,c,a,e) \in \mathcal{E}} l_2((a,e), f(q,d,c))\) where \(l_2\) is some per-example measure.

By extension, one can conceive of a task in which one must also predict a passage \(c\), in addition to an answer and an explanation. One could even integrate QED with a version of the open-domain QA task, which also entails retrieval of documents \(d\). Given QED’s linguistic generality, the data may also be useful as auxiliary input for training models that are not explicitly interested in evaluating explanation generation.

An open question in explainability is how we can build and evaluate models that generate *faithful* explanations, where the explanation truly reflects the model’s underlying reasoning (Jacovi and Goldberg, 2020). Accurate models for the above tasks, even if they do not generate faithful explanations, may still have considerable utility. However, faithful models have several desirable

---

7Passages are the same as NQ long answers.
characteristics (see Sections 5 and 6); we view them as a major avenue for future work.

### 4.2 A SpanBERT Model

The first model we consider for Tasks 1 and 2 uses the SpanBERT coreference resolution model (Joshi et al., 2020; Lee et al., 2017) to identify referential equalities, extends the model with a QA component, and heuristically selects supporting sentences to produce a final QA+QED output.

**Representation** Assume an example contains a question \( q \) of \( m \) tokens \( q_1 \ldots q_m \) and a passage \( c \) consisting of \( n \) tokens \( c_1 \ldots c_n \). We denote the title of the Wikipedia page separately as the sequence \( t \) of \( k \) tokens \( t_1 \ldots t_k \). The model uses SpanBERT to jointly encode the the concatenation of these token sequences,

\[
[\text{CLS}] t_1 \ldots t_k [\text{S1}] q_1 \ldots q_m [\text{S2}] c_1 \ldots c_n [\text{SEP}]
\]

as an input document.8

**Coreference** Given some document \( d \) and a candidate mention \( x \), corresponding to a span within \( d \), define \( \mathcal{Y}(x) \) to be the set of potential antecedents for \( x \). Each antecedent is either a span in the document with start-point before \( x \) in the document, or \( \varepsilon \) signifying that \( x \) does not have an antecedent. We can then define a distribution over the antecedent spans \( \mathcal{Y}(x) \) as \( p(y|x, D) = \frac{\exp(s(x, y))}{\sum_{y' \in \mathcal{Y}(x)} \exp(s(x, y'))} \) where

\[
s(x, y) = \begin{cases} 
0 & \text{if } y = \varepsilon; \\
 s_m(x) + s_m(y) + s_e(x, y) & \text{otherwise}
\end{cases}
\]

\[
s_m(x) = \text{FFNN}_m(g_x)
\]

\[
s_e(x, y) = \text{NN}_c(g_x, g_y)
\]

where \( g_x \) and \( g_y \) are span representations obtained by concatenating the SpanBERT representations of the first and last token in each mention span. The scoring functions \( s_m \) and \( s_e \) represent mention and joint span match scores respectively. Whereas \( s_m \) is a simple feedforward net, \( s_e \) is a more complex scoring function that has been optimized to the coreference task. We refer the reader to Lee et al. (2017) for more details.

Lee et al. (2017) describe a method for training the model based on log-likelihood, and a beam search method that uses the scores \( s_m(\ldots) \) and \( s_e(\ldots) \) to filter candidate mention, antecedent pairs into the final set considered by the loss function. The final output from the coreference model is a hard clustering of the potential mentions into coreference clusters.

Given the constraints of QED referential equalities, we restrict \( s_c \) to only score coreferential links between the query and the passage or between the query and the title (all other values for \( s_m \) or \( s_c \) are set to \(-\infty\)). We model bridges as links between a query passage the title.

We finally post-process the cluster outputs of the coreference component as follows: For each cluster we output the first mention in the cluster that appears in the question with the first mention in the cluster of references that appears in the passage, once cluster mentions are sorted.9 If there is no cluster mention in the passage, we assume the passage reference is a bridge.

**QA** The answer scoring component computes answer candidate representations \( g_z \) using the same candidate mention scoring network as the coreference model, \( \text{FFNN}_m \), as well as a feedforward network, \( \text{FFNN}_a \), that scores candidate answer spans relative to a representation of the question. The score of an answer \( z \) is then computed as

\[
s_a(z) = \text{FFNN}_m(g_z) + \text{FFNN}_a(g_z, g_q).
\]

Thus, the only new parameters belong to a single hidden layer feed-forward net \( \text{FFNN}_a \) that specifically targets the question-answer relationship. Apart from the use of shared candidate mention scoring parameters, no further dependence is introduced between the answer and referential equality predictions.

**Sentence Selection** We perform sentence selection heuristically by choosing the sentence containing the first cluster output by the coreference model. Any subsequent coreference cluster containing a document mention outside of this sentence is dropped in the final prediction. If no referential link is predicted, we take the supporting sentence to be the one containing the answer span.

**Training** For Task 1, we consider an untrained model and a fine-tuned model, both of which omit the QA component described above. In the former, we do not use expert annotated QED data but instead use the CoNLL OntoNotes coreference

---

8We use [S1] = ‘.’ and [S2] = ‘?’ as separators.

9This is necessary because it is technically possible for a cluster to contain more than two mentions before post-processing.
dataset (Pradhan et al., 2012) to train the pretrained SpanBERT model. We only score document mentions in the sentence containing the answer.

For the fine-tuned model, we mark short answers with special tokens before computing the SpanBERT document representation. Then, we further train the model with the training portion of QED data. We used SpanBERT “large”, with a maximum span width of 16 tokens, a top span ratio of 0.2, 30 max antecedents per mention. In fine-tuning, we used an initial learning rate of $3 \cdot 10^{-4}$ and trained for 3 epochs on the QED training set.

For Task 2, we train the QA and Coreference components in a multitask fashion, by minimizing the weighted sum of the QA and coreference cross entropy losses. For the QA data, we augment using passages containing short answers from NQ. Our best results are obtained with a weight of 5 on the coreference loss and 2 epochs of training. The best answer accuracy and QED F1 are obtained for different base learning rates of $2 \cdot 10^{-5}$ and $5 \cdot 10^{-5}$ respectively.

### 4.3 A T5 Model

The second model we consider fine-tunes T5 (Raffel et al., 2020) to predict linearized QA and QED outputs from an input document. We briefly describe the linearization approach here, and refer the reader to Appendix B for a worked example.

**Input Representation** Similar to the SpanBERT model and as depicted immediately above, we pass the concatenation of question, title, and document tokens as input to T5, in that order.$^{10}$ Each input instance is either a QA- or QED-specific instance, which is indicated to T5 by appending a task-specific token to the end of the input.

**Output Representation** The model is tasked with predicting either (1) an answer span or (2) a QED explanation, represented as a sequence of referential equalities, all separated by a special token.$^{11}$ In (2), each referential equality is represented as the concatenation of two spans: the tokens in its query mention and the tokens in its passage mention, separated by ">>". In both (1) and (2) the four tokens in the passage immediately following the answer or passage mention are also appended. These additional tokens are not part of

---

$^{10}$We use ">>" as field separators.

$^{11}$We use ",&&" to separate referential equalities.

---

<table>
<thead>
<tr>
<th>Model</th>
<th>Mention Identification P</th>
<th>R</th>
<th>F1</th>
<th>Mention Alignment P</th>
<th>R</th>
<th>F1</th>
<th>Sentence Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SB-onto</td>
<td>59.0</td>
<td>35.6</td>
<td>44.4</td>
<td>47.7</td>
<td>23.8</td>
<td>35.9</td>
<td>97.3</td>
</tr>
<tr>
<td>SB-fine-tuned</td>
<td>76.8</td>
<td>68.8</td>
<td>72.6</td>
<td>68.4</td>
<td>61.3</td>
<td>64.6</td>
<td>94.2</td>
</tr>
<tr>
<td>T5</td>
<td>73.0</td>
<td>75.8</td>
<td>74.4</td>
<td>63.1</td>
<td>65.5</td>
<td>64.3</td>
<td>95.9</td>
</tr>
</tbody>
</table>

Table 2: SpanBERT (SB) and T5 11B model performance for Task 1: recovering QED annotations when the correct answer is given.

the evaluated spans; they serve to uniquely locate the character offset of the answer or passage mention during evaluation.

Sentence selection proceeds heuristically as in the SpanBERT model.

**Training** We trained T5 11B on only the QED training data, using the standard fine-tuning recipe with a batch size of 1024, learning rate of $2e-4$ and a dropout rate of 0.1. For Task 1, we trained on the explanation task, marking short answers using "«" and "»" brackets in the input. For Task 2, we mixed the QA task and the explanation task with equal weights, and randomly shuffled the instances. We saw the best results when we trained Task 1 for 7000 steps and Task 2 for 2000 steps.

### 4.4 Evaluation and Results

We evaluate answer selection, sentence selection, and the identification of referential equalities. For answer and sentence selection, we report accuracy on 90% span overlap F1. For referential equality, we evaluate both mention identification (the identification of individual referential expressions in the question and passage) and referential equality detection (the identification of pairs of referential expressions).$^{12}$ We compute precision, recall, and F1 measure in both cases.$^{13}$

Results for Task 1 for both the SpanBERT and T5 models are reported in Table 2. The table shows results for both the OntoNotes- and QED-fine-tuned SpanBERT models, as well as the T5 model trained only on the task of explanation prediction. Of note is that trained models trained on QED data do considerably better than the model trained on OntoNotes, indicating that referential

---

$^{12}$Where referential equalities involved bridged passage mentions, we only evaluate the models’ ability to recognize that they are bridged, since there are many conceivable places in a sentence into which mentions can be bridged.

$^{13}$Official evaluation code has been released with the dataset.
Table 3: SpanBERT (SB) and T5 model performance for Task 2: recovering answer and QED annotations given a passage that is known to contain the answer.

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Mention Identification</th>
<th>Mention Alignment</th>
<th>Answer Accuracy</th>
<th>Sentence Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P  R  F1</td>
<td>P  R  F1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SB-QED-only</td>
<td>74.0 63.1 68.1</td>
<td>63.5 54.2 58.6</td>
<td>-</td>
<td>88.4</td>
</tr>
<tr>
<td>SB-QA-only</td>
<td>- - -</td>
<td>- - -</td>
<td>-</td>
<td>73.0</td>
</tr>
<tr>
<td>SB-QA+QED</td>
<td>77.6 64.1 70.4</td>
<td>68.9 57.2 62.5</td>
<td>74.5</td>
<td>90.8</td>
</tr>
<tr>
<td>T5-QED-only</td>
<td>71.1 73.3 72.2</td>
<td>59.5 61.4 60.4</td>
<td>-</td>
<td>88.9</td>
</tr>
<tr>
<td>T5-QA-only</td>
<td>- - -</td>
<td>- - -</td>
<td>-</td>
<td>78.9</td>
</tr>
</tbody>
</table>
| T5-QA+QED             | 70.3 72.3 71.3         | 58.3 59.9 59.1    | 79.2           | 89.1             

In Table 3 we report results for Task 2. SB-QED-only refers to the SpanBERT model fine-tuned only with QED data. SB-QA-only refers to the SpanBERT model fine-tuned on the NQ QA data. SB-QA+QED finetunes on both QA and QED. Similarly, we report results for T5 models. We find that the T5 model tends to have higher recall than the SpanBERT model on mention evaluations, but that the SpanBERT model is considerably more precise. T5 far outperforms SpanBERT on answer accuracy, even though it was fine-tuned without the NQ QA data. Interestingly, in both T5 and SpanBERT models, training on QED data improves QA performance. While the SpanBERT model is more complex than the sequence-to-sequence T5 model, it is considerably more compact (320 million parameters versus 11 billion).

The data contain annotations for answer coreference, in which answer spans outside of the supporting sentence are referred to by an anaphor within it. (See Section 2.) The phenomenon is relatively rare though, and hence there are not enough data to evaluate performance properly. We did perform an additional experiment with T5 where, in addition to an answer span, it predicted its anaphor in the answering sentence where appropriate. The model achieved satisfactory performance, with an F1 of 71%.

5 Rater Study

A major desideratum for explanation generation models is faithfulness—that is, when the explanations generated by a model truly reflect its reasoning process (Jacovi and Goldberg, 2020; Ross et al., 2017). One motivation for this is that when a model is wrong, faithful explanations reliably indicate the reason for the error. In the context of QA, exposing the explanations of a faithful system should improve users’ ability to spot incorrect answers. We show that this is true of a faithful QED system using a rater study.

5.1 Task Setup

Given a question, passage, and a candidate answer span, raters were tasked with assessing whether the candidate answer was correct or incorrect, and indicating the confidence of their assessment. A total of 354 raters, all of whom are US residents and native English speakers, were divided into three disjoint pools to perform the task in three distinct test settings: The None group of raters (n = 121) was presented with a question, passage, and a highlighted answer span. The Sentence group (n = 117) was provided with additional highlighting of the sentence justifying the answer, with no distinction made between referential equalities and predicates. The QED group (n = 116) was provided with additional highlighting to indicate referential equalities between spans in the question and spans in the passage. On average, a given rater provided judgments for 41 questions.

We constructed the data for the study by taking a random set of 50 correct answers, and 50 incorrect guesses from the NQ baseline model (Alberti et al., 2019), on the Natural Questions dev set. So as to ensure that the task was sufficiently challenging, correct instances were the gold answer spans on question/passage pairs where the model produced a false negative—that is, where an answer existed in the passage, but the model was not confident about it. Incorrect instances were false positive guesses from the model, where an answer did not exist in the passage but the model was confident that one did.

Explanations, where present, were manually annotated to simulate the inferences of a hypothetical model that used a QED-style reasoning process. When an item’s answer was correct, the explanation shown was simply its corresponding QED explanation. When the answer was incorrect, referential equalities were identified using counterfactual reasoning; they indicate equivalences that would have to hold if the answer were correct.

Representative examples from the set of rater study items are shown in Figure 5. Note that although referential equalities were manually chosen for incorrect examples, they are not outlandish: They tend to correspond with closely
Figure 5: Three example items from the rater study. In the referential equality error example, the answer is incorrect because the White House and the State Capitol Building are not the same. In the predicate entailment error example, the answer is incorrect because the sentence mentions the number of people over 65, whereas the question asks for the number of people over 50.

related referents, but that are not equivalent upon further inspection. More generally, incorrect answers in the rater study tend to be incorrect for very subtle reasons; this is a result of the aforementioned answer selection process.

All raters were told that highlighting was the output of “an automated question answering system” that was incorrect “about half of the time.” They were advised not to use external knowledge sources or web search to make their judgments. Raters who saw explanations were also told that the system made use of the highlighted explanation to produce its candidate answers.

5.2 Results

Average rater accuracies for each test setting are presented in Table 4. We see that, in aggregate, QED explanations improved accuracy on the task over and above the other test settings, and gave the most improvement on the identification of answers that were incorrect. These improvements translate to incorrect answers resulting from both predicate and reference model errors.

Somewhat surprisingly, highlighting just the sentence containing the answer improved accuracy more than including referential equality highlighting on instances that were correct. This may be because raters’ propensity to mark instances correct decreases as the complexity of explanations increases, from None (73.1%) to Sentence (72.6%) to QED (70.5%).

Also clear from Table 4 is that rater accuracy is much lower on incorrect instances. Even though raters were told that the answers presented were incorrect half of the time, they judged the answers to be correct roughly 71% of the time.14

Figure 6 provides another perspective on the disparity in judgments on correct/incorrect

14While this confirmation bias presents an interesting challenge for future work, it is not a shortcoming of our results: Raters were not trained to do well on the task, as we aimed to approximate how users interact with automated QA systems.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Corr</td>
</tr>
<tr>
<td>None</td>
<td>67.5</td>
<td>90.4</td>
</tr>
<tr>
<td>Sentence</td>
<td>69.7</td>
<td>92.4</td>
</tr>
<tr>
<td>QED</td>
<td>70.2</td>
<td>90.6</td>
</tr>
</tbody>
</table>

Table 4: Rater study results. Corr and Incorr are accuracies of raters in each group on correct and incorrect instances respectively, with incorrect instances further broken into Pred(icate) and Ref(erence) model errors. F1 is on the task of identifying incorrect instances.
instances summarized in Table 4. Highest per-question accuracies in the incorrect pool were still lower than the average accuracy on all correct instances, and the lowest accuracy on incorrect instances is far lower than that of any of the correct instances. The wide distribution of accuracies on incorrect instances ($\sigma \approx 0.50$) seen in Figure 6 was also reflected in the rater pool ($\sigma \approx 0.45$). The challenging nature of incorrect instances speaks to the promise of improvements from QED explanations.

5.3 Effectiveness of explanations

How statistically significant are the results reported in Table 4? The 14,115 test instances were spread across 354 raters and 100 questions. We use the rstanarm R package (Goodrich et al., 2020) to fit a generalized linear mixed model (GLMM) that estimates the log-odds of rater accuracy on the basis of fixed effects (instance correctness and explanation type), while controlling for random effects of rater and question. (See Gelman and Hill (2006) for further discussion of GLMMs.) Ultimately we are interested in the magnitude and statistical properties of the model under the various test settings.

Table 5 shows the fixed effect coefficient and standard deviations for each setting. The presence of QED explanations in the Incorrect setting increased the log-odds of rater accuracy by 0.25, with a posterior predictive p-value of 0.015 that this effect is greater than zero. The comparable effect for Sentence explanations was 0.15, with a posterior predictive p-value of 0.08. The rater and question random effects had standard deviations of 0.63 and 0.90 respectively, reflecting again the high variance of questions shown in Figure 6.

As we saw earlier, the effects of explanations in the Correct setting was reversed: The Sentence explanations caused a small, statistically insignificant increase in log-odds, while QED explanations caused a statistically significant drop in log-odds.

6 Discussion

6.1 QED versus other Explanation Types

QED exists in between relatively unstructured explanation forms on the one hand, such as attention...
Multi-hop

Question: when did martial law end in the philippines

Wikipedia page: Proclamation No. 1081

Passage: Proclamation No 1081 was the proclamation of martial law in the Philippines by President Ferdinand Marcos. It was announced to the public on 23 September 1972, and was formally lifted on 17 January 1981.

Yes/No

Question: can you make and receive calls in airplane mode

Wikipedia page: Airplane_mode

Passage: Airplane mode, aeroplane mode, flight mode, off-air mode, or standby mode, is a setting available on many smartphones, portable computers, and other electronic devices that, when activated, suspends radio-frequency signal transmission by the device, thereby disabling Bluetooth, telephony, and Wi-Fi.

Answer: no

Set-valued

Question: who was involved in the soviet invasion of afghanistan

Wikipedia page: Soviet–Afghan War

Passage: The Soviet–Afghan War lasted over nine years, from December 1979 to February 1989. Insurgent groups known as the mujahideen fought against the Soviet Army and the Democratic Republic of Afghanistan government, mostly in the country’s rural countryside. The mujahideen groups were backed by the United States and Pakistan, making it a Cold War proxy war.

Figure 7: Examples from NQ that go beyond the current definition of QED. Highlighting resembles QED highlighting. In the multi-hop instance, we require an additional sentence to link the entity mentions in the question to the answer sentence. In the yes/no question, the supporting sentence justifies a “No” answer because it contradicts the question predicate. In the set-valued question, multiple sentences provide partial answers to the question, and the resulting answer is the union of all of these.

distributions (Wiegreffe and Pinter, 2019; Jain and Wallace, 2019; Mohankumar et al., 2020) or sequential outputs (Camburu et al., 2018, 2020; Narang et al., 2020; Kumar and Talukdar, 2020) and more elaborate, discrete semantic representations that can in theory be applied to explainable QA (Abzianidze et al., 2017; Wolfson et al., 2020).

6.2 QED and Faithfulness

A major goal for future work is to develop faithful QA models with the QED framework. As Section 5 suggests, models that are not only right for the right reasons, but also wrong for the right reasons, can help users identify subtle errors. Other motivations include model debuggability: Since faithful models should reveal weaknesses in their reasoning, they may enable more targeted intervention.

QED is a promising style of explanation to this end, because it makes use of fundamental semantic variables, like reference (Russell, 1905; Clark and Marshall, 1981; Tomasello et al., 2007). We can say, definitively, that in order for a sentence to answer a question about a thing, its meaning must involve that thing in a very particular sense. Posed counterfactually, when you break referential equality, you break answerhood, and the same argument follows for predicate entailment. This is a hallmark of a good explanation (Pearl, 2019; Lipton, 2001).

6.3 Scoping and Extension to other Question Types

The instantiation of QED presented in the current work is limited to extractive wh-questions whose answers are entailed by single sentences. We feel this scoping is well justified, because (1) a significant portion of NQ falls under QED’s current purview; (2) previous work and data analysis suggests QED can be readily extended to accommodate these other types (Hearst, 1992; Miltsakaki et al., 2004; Lamm et al., 2018; Tandon et al., 2019); and (3) close study of the single sentence case is a necessary condition for these other question types.

In Figure 7, we present several representative NQ instances that require more machinery than QED provides at present. Let us consider how QED might be extended to handle each of these.
Multi-hop QA For multi-hop questions, referential equalities involve longer, text-mediated paths from entity references in the question to an ultimate sentence entailing its answer (Yang et al., 2018).

Yes/No QA Answering Yes/No questions requires identifying sentences in a text that entail or contradict the premise presented in the question.

Set-valued QA Set-valued QA requires assembling QED explanations for a set of answers to the question, and returning the union of the (unique) answers found.

Looking further afield from these question types above, which are less frequent in NQ but nevertheless attested there, it becomes clear that QA writ large is much broader than even a dataset of NQ’s scale suggests (Rogers et al., 2020). The generality of QED as a model for how elements of questions can link up with textual evidence suggests that QED would likely be complementary to, rather than at odds with, efforts to understand these broader senses of QA.

7 Conclusions

We have described QED, a framework for explanations in question answering, and we have introduced a dataset of QED annotations. The framework is grounded in referential equality, and entailment. In addition we have described baseline models for two QED-based tasks, and a rater study utilizing QED annotations.

Future work should consider the development of models based on QED, especially those that provide faithful explanations, and extensions of QED beyond the single-sentence assumption.

References


We now describe the form of the e and a annotations, making reference to the following example. Subscripts indicate token positions:

**Question:** whom won wimbledon in 2019

**Passage:** Simona Halep is a female tennis player. She won Wimbledon in 2019.

As a preliminary step, given the paragraph c and sentence s, we use S to refer to the set of all phrases within s. Our initial definition of S is

\[ S = \{(i, j) : s_0 \leq i \leq j \leq s_1\} \]

We also define the set of question phrases Q and passage phrases C to be

\[ Q = \{(i, j) : 1 \leq i \leq j \leq m\} \]

\[ C = \{(i, j) : 1 \leq i \leq j \leq n\} \]

We can then give the following definitions:

**Definition 1** Each referential equality annotation \( e_k \) for \( k = 1 \ldots |e| \) is a pair \((\phi_k, \pi_k) \in Q \times S\), specifying that the phrase \( \phi_k \) in the query refers to the same thing in the world as the phrase \( \pi_k \) within \( s \).

In our example,

\[ e = \{(3,3), (11,11), (5,5), (13,13)\} \]

where the first tuple in the sequence corresponds with the alignment between “wimbledon” in the question and “Wimbledon” in the passage, and the second tuple with “2019” in the question and “2019” in the passage.

**Definition 2** Each answer annotation \( a_k \) for \( k = 1 \ldots |a| \) is a pair \((\pi_k, \xi_k) \in S \times C\) specifying that the answer is given by phrase \( \pi_k \), and the full string corresponding to \( \pi_k \) after coreference resolution is the phrase \( \xi_k \). If no coreference resolution is required then \( \pi_k = \xi_k \).

In our example,

\[ a = \{(9,9), (1,2)\} \]

corresponding with the alignment of “She” in the sentence “She won Wimbledon in 2019” with the mention of “Simona Halep” earlier in the passage.
A.1 Extending Annotations to Include Bridging

Recall the definition of bridging in Section 2. We extend the formal definition of QED to include bridging by redefining $S$ to include implicit phrases introduced in the form of implicit prepositional phrases, as in the “winner [of ...]”. The modified definition of $S$ includes all phrases of the following form: (1) Any pair $(i,j)$ such that $s_0 \leq i \leq j \leq s_1$ indicating the subsequence of words $c_i \ldots c_j$ within the sentence. (2) Any triple $(i,j,p)$ such that $s_0 \leq i \leq j \leq s_1$ and $p$ is a preposition, indicating the implicit noun phrase in the sentence that modifies the phrase $c_i \ldots c_j$ through the preposition $p$. (3) Any pair $\text{NULL},p$ such that $p$ is a preposition, indicating the implicit noun phrase modifying the entire sentence $c_{s_0} \ldots c_{s_1}$ through the preposition $p$.

Given the following example, then:

**Question:** who1 won2 america’s3 got4 talent5 season6 117

**Passage:** The1 11th2 season3 of4 America’s5 Got6 Talent7 began8 broadcasting9 in10 the11 United12 States13 during14 201615 -16 Grace17 VanderWaal18 was19 announced20 as21 the22 winner23 on24 September25 1426 -27 201628 -29

we have that

$$e = [((3,7),(22,23),”of”)]
$$

This means that the question span “america’s got talent season 11” is bridged by the reference “the winner” in the answering sentence. The preposition “of” indicates how the question referent can be attached to the sentence reference: Putting them together yields “the winner of america’s got talent season 11”.

B T5 Model Linearization

We describe our method for linearizing QED instances as T5 input and output sequences. Let us consider the following example:

**Question:** how many seats in university of michigan stadium

**Passage:** Michigan Stadium, nicknamed “The Big House”, is the football stadium for the University of Michigan in Ann Arbor, Michigan. It is the largest stadium in the United States and the second largest stadium in the world. Its official capacity is 107,601.

Recall that the QED annotation for this example is as follows

$$\text{how many seats in [=1 university of michigan stadium]}$$

$$[[=1 \text{ Its}] \text{ official capacity is } [=A 107,601]]$$

The T5 model input are constructed by concatenating the question, page title, and paragraph into one sequence, separated by “>>”:

how many seats in university of michigan stadium >> Michigan Stadium >> Michigan Stadium, nicknamed “The Big House”, is the football stadium for the University of Michigan in Ann Arbor, Michigan. It is the largest stadium in the United States and the second largest stadium in the world. Its official capacity is 107,601.

A task-specific token is prepended to this input to indicate whether the model should produce an answer or an explanation. The answer output sequence is as follows:

107,601 ».

where additional material after the special character ”>” (as distinct from “>>”) is used to disambiguate the position of the answer in the passage. Finally, the explanation would be linearized as follows:

university of michigan stadium >> Its » official capacity is 107,601

Here, the first phrase corresponds to the question mention and the second to the passage mention. The additional material after the passage mention is meant to uniquely identify its position in the passage, for evaluation purposes.