BREAK It Down: A Question Understanding Benchmark

Tomer Wolfson\textsuperscript{1,3}, Mor Geva\textsuperscript{1,3}, Ankit Gupta\textsuperscript{1}, Matt Gardner\textsuperscript{3}, Yoav Goldberg\textsuperscript{2,3}, Daniel Deutch\textsuperscript{1}, Jonathan Berant\textsuperscript{1,3}

\textsuperscript{1}Tel Aviv University \hspace{1em} \textsuperscript{2}Bar-Ilan University \hspace{1em} \textsuperscript{3}Allen Institute for AI
{tomerwol, morgeva}@mail.tau.ac.il, ankitgupta.iitkanpur@gmail.com, mattg@allenai.org, yoav.goldberg@gmail.com, daniellede@post.tau.ac.il, joberant@cs.tau.ac.il

Abstract

Understanding natural language questions entails the ability to break down a question into the requisite steps for computing its answer. In this work, we introduce a Question Decomposition Meaning Representation (QDMR) for questions. QDMR constitutes the ordered list of steps, expressed through natural language, that are necessary for answering a question. We develop a crowdsourcing pipeline, showing that quality QDMRs can be annotated at scale, and release the BREAK dataset, containing over 83K pairs of questions and their QDMRs. We demonstrate the utility of QDMR by showing that (a) it can be used to improve open-domain question answering on the HotPotQA dataset, (b) it can be deterministically converted to a pseudo-SQL formal language, which can alleviate annotation in semantic parsing applications. Last, we use BREAK to train a sequence-to-sequence model with copying that parses questions into QDMR structures, and show that it substantially outperforms several natural baselines.

1 Introduction

Recently, increasing work has been devoted to models that can reason and integrate information from multiple parts of an input. This includes reasoning over images (Antol et al., 2015; Johnson et al., 2017; Suhr et al., 2019; Hudson and Manning, 2019), paragraphs (Dua et al., 2019), documents (Welbl et al., 2018; Talmor and Berant, 2018; Yang et al., 2018), tables (Pasupat and Liang, 2015), and more. Question answering (QA) is commonly used to test the ability to reason, where a complex natural language question is posed, and is to be answered given a particular context (text, image, etc.). Although questions often share structure across tasks and modalities, understanding the language of complex questions has thus far been addressed within each task in isolation. Consider the questions in Figure 1, all of which express operations such as fact chaining and counting. Additionally, humans can take a complex question and break it down into a sequence of simpler questions even when they are unaware of what or where the answer is. This ability, to compose and decompose questions, lies at the heart of human language (Pelletier, 1994) and allows us to tackle previously unseen problems. Thus, better question understanding models should improve performance and generalization in tasks that require multi-step reasoning or that do not have access to substantial amounts of data.

In this work we propose question understanding as a standalone language understanding task. We introduce a formalism for representing the meaning of questions that relies on question decomposition, and is agnostic to the information source. Our formalism, Question Decomposition Meaning Representation (QDMR), is inspired by database query languages (SQL; SPARQL), and by semantic parsing (Zelle and Mooney, 1996; Zettlemoyer and Collins, 2005; Clarke et al., 2010), in which questions are given full meaning representations.

We express complex questions via simple ("atomic") questions that can be executed in sequence to answer the original question. Each atomic question can be mapped into a small set of operations, where each operation either selects a set of entities, retrieves information about
their attributes, or aggregates information over entities. While this has been formalized in knowledge-base (KB) query languages (Chamberlin and Boyce, 1974), the same intuition can be applied to other modalities, such as images and text. QDMR abstracts away the context needed to answer the question, allowing in principle to query multiple sources for the same question.

In contrast to semantic parsing, QDMR operations are expressed through natural language, facilitating annotation at scale by non-experts. Figure 1 presents examples of complex questions on three different modalities. The middle box lists the natural language decompositions provided for each question, and the bottom box displays their corresponding formal queries.

QDMR serves as the formalism for creating Break, a question decomposition dataset of 83,978 questions over ten datasets and three modalities. Break is collected via crowdsourcing, with a user interface that allows us to train crowd-workers to produce quality decompositions (§3). Validating the quality of annotated structures reveals 97.4% to be correct (§4).

We demonstrate the utility of QDMR in two setups. First, we regard the task of open-domain QA over multi-hop questions from the HotpotQA dataset. Combining QDMR structures in Break with a reading comprehension (RC) model (Min et al., 2019b) improves $F_1$ from 43.3 to 52.4 (§5). Second, we show that decompositions in Break possess high annotation consistency, which indicates that annotators produce high-quality QDMRs (§4.3). In §6 we discuss how these QDMRs can be used as a strong proxy for full logical forms in semantic parsing.

We use Break to train a neural QDMR parser that maps questions into QDMR representations, based on a sequence-to-sequence model with copying (Gu et al., 2016). Manual analysis of generated structures reveals an accuracy of 54%, showing that automatic QDMR parsing is possible, though still far from human performance (§7).

To conclude, our contributions are:

- Proposing the task of question understanding and introducing the QDMR formalism for representing the meaning of questions (§2)
- The Break dataset, which consists of 83,978 examples sampled from 10 datasets over three distinct information sources (§3)
- Showing how QDMR can be used to improve open-domain question answering (§5), as well as alleviate the burden of annotating logical forms in semantic parsing (§6)
- A QDMR parser based on a sequence-to-sequence model with copying mechanism (§7)

The Break dataset, models, and entire codebase are publicly available at: https://github.com/tomerwol/g/Break.

2 Question Decomposition Formalism

In this section we define the QDMR formalism for domain agnostic question decomposition.

QDMR is primarily inspired by SQL (Codd, 1970; Chamberlin and Boyce, 1974). However, while SQL was designed for relational databases, QDMR also aims to capture the meaning of questions over unstructured sources such as text and images. Thus, our formalism abstracts away from SQL by assuming an underlying “idealized” KB, which contains all entities and relations expressed in the question. This abstraction enables QDMR to be unrestricted to a particular modality, with its operators to be executed also against text.
<table>
<thead>
<tr>
<th>Operator</th>
<th>Template / Signature</th>
<th>Question</th>
<th>Decomposition</th>
</tr>
</thead>
</table>
| Select   | Return [entities]    | How many touchdowns were scored overall? | 1. Return touchdowns  
2. Return the number of #1 |
| Filter   | Return [ref] [condition] | I would like a flight from Toronto to San Diego please. | 1. Return flights  
2. Return #1 from Toronto  
3. Return #2 to San Diego |
2. Return the head coach of #1 |
2. Return border states of #1  
3. Return the number of #2 |
| Group    | Return [aggregate] [ref1] for each [ref2] | How many female students are there in each club? | 1. Return clubs  
2. Return female students of #1  
3. Return the number of #2 for each #1 |
| Superlative | Return [ref1] where [ref2] is [highest / lowest] | What is the keyword, which has been contained by the most number of papers? | 1. Return papers  
2. Return keywords of #1  
3. Return the number of #1 for each #2  
4. Return #2 where #3 is highest |
| Comparative | Return [ref1] where [ref2] [comparison] [number] | Who are the authors who have more than 500 papers? | 1. Return authors  
2. Return papers of #1  
3. Return the number of #2 for each of #1  
4. Return #1 where #3 is more than 500 |
| Union    | Return [ref1], [ref2] | Tell me who the president and vice-president are? | 1. Return the president  
2. Return the vice-president  
3. Return #1, #2 |
| Intersection | Return [relation] in both [ref1] and [ref2] | Show the parties that have representatives in both New York state and representatives in Pennsylvania state. | 1. Return representatives  
2. Return #1 in New York state  
3. Return #1 in Pennsylvania state  
4. Return parties in both #2 and #3 |
| Discard  | Return [ref1] besides [ref2] | Find the professors who are not playing Canoeing. | 1. Return professors  
2. Return #1 playing Canoeing  
3. Return #1 besides #2 |
| Sort     | Return [ref1] sorted by [ref2] | Find all information about student addresses, and sort by monthly rental. | 1. Return students  
2. Return addresses of #1  
3. Return number of #1  
4. Return #2 sorted by #3 |
| Boolean  | Return [if / is] [ref1] [condition] | Were Scott Derrickson and Ed Wood of the same nationality? | ...  
3. Return the nationality of #1  
4. Return the nationality of #2  
5. Return if #3 is the same as #4 |
| Arithmetic | Return the [arithmetic] of [ref1] and [ref2] | How many more red objects are there than blue objects? | ...  
3. Return the number of #1  
4. Return the number of #2  
5. Return the difference of #3 and #4 |

<table>
<thead>
<tr>
<th>Operator</th>
<th>Template / Signature</th>
<th>Question</th>
<th>Decomposition</th>
</tr>
</thead>
</table>
| Arithmetic | Return the [arithmetic] of [ref1] and [ref2] | How many more red objects are there than blue objects? | ...  
3. Return the number of #1  
4. Return the number of #2  
5. Return the difference of #3 and #4 |

Table 1: The 13 operator types of QDMR steps. Listed are, the natural language template used to express the operator, the operator signature, and an example question that uses the query operator in its decomposition.

and images, while allowing in principle to query multiple modalities for the same question.\(^1\)

**QDMR Definition** Given a question \(x\), its QDMR is a sequence of \(n\) steps, \(s = (s^1, \ldots, s^n)\), where each step \(s^i\) corresponds to a single query operator \(f^i\) (see Table 1). A step, \(s^i\) is a sequence of tokens, \(s^i = (s^i_1, \ldots, s^i_{n_i})\), where a token \(s^i_k\) is either a word from a predefined lexicon \(L_x\) (details in \(\S 3\)) or a reference token, referring to the result of a previous step \(s^j\), where \(j < i\). The last step, \(s^n\) returns the answer to \(x\).

**Decomposition Graph** QDMR structures can be represented as a directed acyclic graph (DAG), used for evaluating QDMR parsing models (\(\S 7.1\)).

\(^1\)A system could potentially answer “Name the political parties of the most densely populated country”, by retrieving “the most densely populated country” using a database query, and “the political parties of #1” via an RC model.
Given QDMR, \( s = \langle s^1, ..., s^n \rangle \), each step \( s^i \) is a node in the graph, labeled by its sequence of tokens and index \( i \). Edges in the graph are induced by reference tokens to previous steps. Node \( s^i \) is connected by an incoming edge \( \langle s^j, s^i \rangle \), if \( \text{ref}[s^j] \in \{ s^i \}, \ldots, s^k \rangle \). That is, if one of the tokens in \( s^i \) is a reference to \( s_j \), Figure 2 displays a sequence of QDMR steps, represented as a DAG.

**QDMR Operators** A QDMR step corresponds to one of 13 query operators. We designed the operators to be expressive enough to represent the meaning of questions from a diverse set of datasets (§3). QDMR assumes an underlying KB, \( K \), which contains all of the entities and relations expressed in its steps. A relation, \( r \), is a function mapping two arguments to whether \( r \) holds in \( K \): \( \langle r(x, y) \rangle \in \{ \text{true, false} \} \). The operators operate over: (i) sets of objects \( S_o \), where objects \( o \), are either numbers \( n \), boolean values \( b \), or entities \( e \) in \( K \); (ii) a closed set of phrases \( w_{op} \), describing logical operations; and (iii) natural language phrases \( w \), representing entities and relations in \( K \). We assume the existence of grounding functions that map a phrase \( w \) to concrete constants in \( K \). Table 2 describes the aforementioned constructs. In addition, we define the function \( \text{map}_{K}(S_e, S_o) \) which maps entity \( e \in S_e \) to the set of corresponding objects from \( S_o \). Each \( o \in S_o \) corresponds to an \( e \in S_e \) by being contained in the result of a sequence of PROJECT and GROUP operations applied to \( e \):

\[
\text{map}_{K}(S_e, S_o) = \{ \langle e, o \rangle \mid e \in S_e, o \in S_o, o \in \text{op}_k \circ \ldots \circ \text{op}_1(e) \}.
\]

We now formally define each QDMR operator and provide concrete examples in Table 1.

- **SELECT**: Computes the set of entities in \( K \) corresponding to \( w \): \( \text{select}(w) = \text{ground}_{K}^{e}(w) \).

---

Given QDMR, \( \langle s^1, ..., s^n \rangle \), each step \( s^i \) is a node in the graph, labeled by its sequence of tokens and index \( i \). Edges in the graph are induced by reference tokens to previous steps. Node \( s^i \) is connected by an incoming edge \( \langle s^j, s^i \rangle \), if \( \text{ref}[s^j] \in \{ s^i \}, \ldots, s^k \rangle \). That is, if one of the tokens in \( s^i \) is a reference to \( s_j \), Figure 2 displays a sequence of QDMR steps, represented as a DAG.

**QDMR Operators** A QDMR step corresponds to one of 13 query operators. We designed the operators to be expressive enough to represent the meaning of questions from a diverse set of datasets (§3). QDMR assumes an underlying KB, \( K \), which contains all of the entities and relations expressed in its steps. A relation, \( r \), is a function mapping two arguments to whether \( r \) holds in \( K \): \( \langle r(x, y) \rangle \in \{ \text{true, false} \} \). The operators operate over: (i) sets of objects \( S_o \), where objects \( o \), are either numbers \( n \), boolean values \( b \), or entities \( e \) in \( K \); (ii) a closed set of phrases \( w_{op} \), describing logical operations; and (iii) natural language phrases \( w \), representing entities and relations in \( K \). We assume the existence of grounding functions that map a phrase \( w \) to concrete constants in \( K \). Table 2 describes the aforementioned constructs. In addition, we define the function \( \text{map}_{K}(S_e, S_o) \) which maps entity \( e \in S_e \) to the set of corresponding objects from \( S_o \). Each \( o \in S_o \) corresponds to an \( e \in S_e \) by being contained in the result of a sequence of PROJECT and GROUP operations applied to \( e \):

\[
\text{map}_{K}(S_e, S_o) = \{ \langle e, o \rangle \mid e \in S_e, o \in S_o, o \in \text{op}_k \circ \ldots \circ \text{op}_1(e) \}.
\]

We now formally define each QDMR operator and provide concrete examples in Table 1.

- **SELECT**: Computes the set of entities in \( K \) corresponding to \( w \): \( \text{select}(w) = \text{ground}_{K}^{e}(w) \).

---

The sequence of operations \( \text{op}_1, \ldots, \text{op}_k \) is traced using the references to previous steps in the QDMR structure.
• COMPARATIVE: Receives entity set $S_e$ and number set $S_n$. Each $n \in S_n$ is the result of a mapping from an $e \in S_e$. It returns a subset of $S_e$ for which the comparison with $n'$, represented by $w_{com}$, holds.

$$com(n, n') \equiv true.$$  

comparative($S_e, S_n, w_{com}, n') = \{ e \mid \langle e, n \rangle \in \text{map}_K(S_e, S_n), com(n, n') \equiv true \}.$$

• UNION: Denotes the union of object sets:

$$\text{union}(S^1, S^2) = S^1 \cup S^2.$$  

• DISCARD: Denotes the set difference of two object sets:

$$\text{discard}(S^1, S^2) = S^1 \setminus S^2.$$  

• INTERSECTION: Computes the intersection of its entity sets and returns all objects which relate to the entities with the relation expressed by $w$.

$$\text{intersect}(w, S^1, S^2) = \{ o \mid e \in S^1 \cap S^2, [r(e, o)]_K \equiv true, r = \text{ground}_K(w) \}.$$  

• SORT: Orders a set of entities according to a corresponding set of numbers. Each number $n_i$ is the result of a mapping from entity $e_i$.

$$\text{sort}(S_e, S_n) = \{ \langle e_i \ldots e_i \rangle \mid \langle e_i, n_i \rangle \in \text{map}_K(S_e, S_n), n_i \leq \ldots \leq n_{im} \}.$$  

• BOOLEAN: Returns whether the relation expressed by $w$ holds between the input objects:

$$\text{boolean}(S^1, S^2, w) = \{ [r(o_1, o_2)]_K \}, \text{where } r = \text{ground}_K(w) \text{ and } S^1, S^2 \text{ are singleton sets containing } o_1, o_2 \text{ respectively.}$$  

• ARITHMETIC: Computes the application of an arithmetic operation:

$$\text{arith}(w_{arith}, S^1, S^2) = \{ \text{ari}(n_1, n_2) \}, \text{where } S^1, S^2 \text{ are singleton sets containing } n_1, n_2 \text{ respectively.}$$  

**High-level Decompositions** In QDMR, each step corresponds to a single logical operator. In certain contexts, a less granular decomposition might be desirable, where sub-structures containing multiple operators could be collapsed to a single node. This can be easily achieved in QDMR by merging certain adjacent nodes in its DAG structure. When examining existing RC datasets (Yang et al., 2018; Dua et al., 2019), we observed that long spans in the question often match long spans in the text, due to existing practices of generating questions via crowdsourcing. In such cases, decomposing the long spans into multiple steps and having an RC model process each step independently, increases the probability of error. Thus, to promote the usefulness of QDMR for current RC datasets and models, we introduce high-level QDMR, by merging the following operators:

- **SELECT + PROJECT on named entities:** For the question, “What is the birthdate of Jane?” its high-level QDMR would be “return the birthdate of Jane” as opposed to the more granular, “return Jane; return birthdate of #1”.

- **SELECT + FILTER:** Consider the first step of the example in Figure 3. It contains both a SELECT operator (“return actress”) as well as two FILTER conditions (“return actress” and “on the TV sitcom…”).

- **FILTER + GROUP + COMPARATIVE:** Certain high-level FILTER steps contain implicit grouping and comparison operations. E.g., “return yard line scores in the fourth quarter; return #1 that both teams scored from”. Step #2 contains an implicit GROUP of team per yard line and a COMPARATIVE returning the lines where exactly two teams scored.

We provide both granular and high-level QDMRs for a random subset of RC questions (see Table 3). The concrete utility of high-level QDMR to open-domain QA is presented in §5.

### 3 Data Collection

Our annotation pipeline for generating Break consisted of three phases. First, we collected complex questions from existing QA benchmarks. Second, we crowdsourced the QDMR annotation of these questions. Finally, we validated worker annotations in order to maintain their quality.
Table 3: The QA datasets in Break. Lists the number of examples in the original dataset and in Break. Numbers of high-level QDMRs are denoted by high.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Example</th>
<th>Original</th>
<th>Break</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aiming</td>
<td>Return me the total citations of all the papers in the VLDB conference.</td>
<td>195</td>
<td>195</td>
</tr>
<tr>
<td>Atlanta</td>
<td>What is the first flight from Atlanta to Baltimore that serves lunch?</td>
<td>5,283</td>
<td>9,060</td>
</tr>
<tr>
<td>GeoTrivia</td>
<td>How long is the highest point in the browse area?</td>
<td>888</td>
<td>877</td>
</tr>
<tr>
<td>Speech</td>
<td>How many minutes correspond to each invoice number?</td>
<td>10,181</td>
<td>7,982</td>
</tr>
<tr>
<td>Visual-Humans</td>
<td>What is the number of cylinders divided by the number of cakes?</td>
<td>32,164</td>
<td>13,935</td>
</tr>
<tr>
<td>NUSQL</td>
<td>If there are only two dogs pulling one of the sleds?</td>
<td>28,680</td>
<td>13,517</td>
</tr>
<tr>
<td>CoQA</td>
<td>What was Gandhi's occupation before becoming a freedom fighter?</td>
<td>11,214</td>
<td>5,520</td>
</tr>
<tr>
<td>CWQ</td>
<td>Robert &amp; Jordan is part of the organization started by whom?</td>
<td>34,689</td>
<td>2,991</td>
</tr>
<tr>
<td>DROP</td>
<td>Approximately how many years did the churches built in 1909 survive?</td>
<td>96,567</td>
<td>10,262</td>
</tr>
<tr>
<td>Hard-CQA</td>
<td>Benjamin Halpert was a footballer for a club that plays in home matches where?</td>
<td>23,066</td>
<td>10,181</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>83,978</td>
<td></td>
</tr>
</tbody>
</table>

Question Collection Questions in Break were randomly sampled from ten QA datasets over the following tasks (Table 3):

- **Semantic Parsing**: Mapping natural language utterances into formal queries, to be executed on a target KB (Price, 1990; Zelle and Mooney, 1996; Li and Jagadish, 2014; Yu et al., 2018).

- **Reading Comprehension (RC)**: Questions that require understanding of a text passage by reasoning over multiple sentences (Talmor and Berant, 2018; Yang et al., 2018; Dua et al., 2019; Abujabal et al., 2019).

- **Visual Question Answering (VQA)**: Questions over images that require both visual and numerical reasoning skills (Johnson et al., 2017; Suhr et al., 2019).

All questions collected were composed by human annotators.\(^3\) HopHopQA questions were all sampled from the hard split of the dataset.

QDMR Annotation A key question is whether it is possible to train non-expert annotators to produce high-quality QDMRs. We designed an annotation interface (Figure 4), where workers are first given explanations and examples on how to identify and phrase each of the operators in Table 1. Then, workers decompose questions into a list of steps, where they are only allowed to use words from a lexicon \(L_x\), which contains: (a) words appearing in the question (or their automatically computed inflections), (b) words from a small pre-defined list of 66 function word such as, ‘if’, ‘on’, ‘for each’, or (c) reference tokens that refer to the results of a previous step. This ensures that the language used by workers is consistent across examples, while being expressive enough for the decomposition. Our annotation interface presents workers with the question only, so they are agnostic to the original modality of the question. The efficacy of this process is explored in §4.2.

We used Amazon Mechanical Turk to crowdsource QDMR annotation. In each task, workers decomposed questions, paying them $0.40 per question, which amounts to an average pay of $12 per hour. Overall, we collected 83,978 examples using 64 distinct workers. The dataset was partitioned into train/development/test sets following the partitions in the original datasets. During partition, we made sure that development and test samples do not share the same context.

Worker Validation To ensure worker quality, we initially published qualification tasks, open to all workers in the United States. The task required workers to carefully review the annotation instructions and decompose 10 example questions. The examples were selected so that each QDMR operation should appear in at least one of their decompositions (Table 1). In total, 64 workers were able to correctly decompose at least 8 examples and were qualified as annotators. To validate worker performance over time, we conducted random validations of annotations. Over 9K annotations were

\(^3\)Except for ComplexWebQuestions (CWQ), where annotators paraphrased automatically generated questions.
Table 4: Operator prevalence in Break. Lists the percentage of QDMRs where the operator appears.

<table>
<thead>
<tr>
<th>Operator</th>
<th>QDMR</th>
<th>QDMR&lt;sub&gt;high&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>SELECT</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>PROJECT</td>
<td>69.0%</td>
<td>35.6%</td>
</tr>
<tr>
<td>FILTER</td>
<td>53.2%</td>
<td>15.3%</td>
</tr>
<tr>
<td>AGGREGATE</td>
<td>38.1%</td>
<td>22.3%</td>
</tr>
<tr>
<td>BOOLEAN</td>
<td>30.0%</td>
<td>4.6%</td>
</tr>
<tr>
<td>COMPARATIVE</td>
<td>17.0%</td>
<td>1.0%</td>
</tr>
<tr>
<td>GROUP</td>
<td>9.7%</td>
<td>0.7%</td>
</tr>
<tr>
<td>SUPERLATIVE</td>
<td>6.3%</td>
<td>13.0%</td>
</tr>
<tr>
<td>UNION</td>
<td>5.5%</td>
<td>0.5%</td>
</tr>
<tr>
<td>ARITHMETIC</td>
<td>5.4%</td>
<td>11.2%</td>
</tr>
<tr>
<td>DISCARD</td>
<td>3.2%</td>
<td>1.2%</td>
</tr>
<tr>
<td>INTERSECTION</td>
<td>2.7%</td>
<td>2.8%</td>
</tr>
<tr>
<td>SORT</td>
<td>0.9%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Total</td>
<td>60,150</td>
<td>23,828</td>
</tr>
</tbody>
</table>

Table 5: The distribution over QDMR sequence length.

<table>
<thead>
<tr>
<th>Steps</th>
<th>QDMR</th>
<th>QDMR&lt;sub&gt;high&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–2</td>
<td>10.7%</td>
<td>59.8%</td>
</tr>
<tr>
<td>3–4</td>
<td>44.9%</td>
<td>31.6%</td>
</tr>
<tr>
<td>5–6</td>
<td>27.0%</td>
<td>7.9%</td>
</tr>
<tr>
<td>7–8</td>
<td>10.1%</td>
<td>0.6%</td>
</tr>
<tr>
<td>9+</td>
<td>7.4%</td>
<td>0.2%</td>
</tr>
</tbody>
</table>

4 Regarding the three merged operators of high-level QDMRs (§2), the first two operators are treated as SELECT, while the third is considered a FILTER.

4.2 Quality Analysis

We describe the process of estimating the correctness of collected QDMR annotations. Similar to previous works (Yu et al., 2018; Kwiatkowski et al., 2019) we use expert judgments, where the experts had prepared the guidelines for the annotation task. Given a question and its annotated QDMR, (q,s) the expert determines the correctness of s using one of the following categories:

- Correct (C): If s constitutes a list of QDMR operations that lead to correctly answering q.
- Granular (C<sub>G</sub>): If s is correct and none of its operators can be further decomposed.<sup>5</sup>
- Incorrect (I): If s is in neither C nor C<sub>G</sub>.

Examples of these expert judgments are shown in Figure 5. To estimate expert judgment of correctness, we manually reviewed a random sample of 500 QDMRs from Break. We classified 93.8% of the samples in C<sub>G</sub> and another 3.6% in C. Thus, 97.4% of the samples constitute a correct decomposition of the original question. Workers have somewhat struggled with decomposing superlatives (e.g., ‘biggest sphere’), as evident from the first question in Figure 5. Collected QDMRs displayed similar estimates of C, C<sub>G</sub>, and I, regardless of their modality (DB, text, or image).

4.3 Annotation Consistency

As QDMR is expressed using natural language, it introduces variability into its annotations. We wish to validate the consistency of collected QDMRs, that is, whether we can correctly infer the formal

5For high-level QDMRs, the merged operators (§2) are considered to be fully decomposed.
Figure 5: Examples and justifications of expert judgment on collected QDMRs in Break.

QDMR operator \( f^i \) and its arguments from each step \( s^i \). To infer these formal representations, we developed an algorithm that goes over the QDMR structure step-by-step, and for each step \( s^i \), uses a set of predefined templates to identify \( f^i \) and its arguments, expressed in \( s^i \). This results in an execution graph (Figure 2), where the execution result of a parent node serves as input to its child.

Overall, 99.5% of QDMRs had all their steps mapped into pseudo-logical forms by our algorithm. To evaluate the correctness of the mapping algorithm, we randomly sampled 350 logical forms, and examined the structure of the formulas, assuming that words copied from the question correspond to entities and relations in an idealized KB (see §2). Of this sample, 99.4% of its examples had all of their steps, \( s^i \), correctly mapped to the corresponding \( f^i \). Overall, 93.1% of the examples were of fully accurate logical forms, with errors being due to QDMRs that were either incorrect or not fully decomposed (\( I, C \) in §4.2). Thus, a rule-based algorithm can map more than 93% of the annotations into a correct formal representation. This shows that our annotators produced consistent and high-quality QDMRs. Moreover, it suggests that non-experts can annotate questions with pseudo-logical forms, which can be used as a cheap intermediate representation for semantic parsers (Yih et al., 2016), further discussed in §6.

5 QDMR for Open-domain QA

A natural setup for QDMR is in answering complex questions that require multiple reasoning steps. We compare models that exploit question decompositions to baselines that do not. We use the open-domain QA (“full-wiki”) setting of the HotpotQA dataset (Yang et al., 2018): Given a question, the QA model retrieves the relevant Wikipedia paragraphs and answers the question using these paragraphs.

5.1 Experimental Setup

We compare BreakRC, a model that utilizes question decomposition to BERTQA, a standard QA model, based on BERT (Devlin et al., 2019), and present Combined, an approach that enjoys the benefits of both models.

**Algorithm 1 BreakRC**

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>procedure BreakRC(( \text{QDMR} ))</td>
</tr>
<tr>
<td>2.</td>
<td>( \text{ansrs} \leftarrow \emptyset )</td>
</tr>
<tr>
<td>3.</td>
<td>for ( s^i ) in ( s = {s^1, \ldots, s^n} ) do</td>
</tr>
<tr>
<td>4.</td>
<td>( \text{op} \leftarrow \text{OpType}(s^i) )</td>
</tr>
<tr>
<td>5.</td>
<td>( \text{refs} \leftarrow \text{ReferencedSteps}(s^i) )</td>
</tr>
<tr>
<td>6.</td>
<td>if ( \text{op} ) is ( \text{SELECT} ) then</td>
</tr>
<tr>
<td>7.</td>
<td>( \text{ans} \leftarrow \text{Answer}(s^i) )</td>
</tr>
<tr>
<td>8.</td>
<td>else if ( \text{op} ) is ( \text{FILTER} ) then</td>
</tr>
<tr>
<td>9.</td>
<td>( s^i \leftarrow \text{ExtractQuestion}(s^i) )</td>
</tr>
<tr>
<td>10.</td>
<td>( \text{ans}_{\text{sup}} \leftarrow \text{Answer}(s^i) )</td>
</tr>
<tr>
<td>11.</td>
<td>( \text{ans} \leftarrow \text{Answer}(\text{refs}[0]) )</td>
</tr>
<tr>
<td>12.</td>
<td>else if ( \text{op} ) is ( \text{COMPARISON} ) then</td>
</tr>
<tr>
<td>13.</td>
<td>( \text{ans} \leftarrow \text{CompareSteps}(\text{refs}, s^i) )</td>
</tr>
<tr>
<td>14.</td>
<td>else do ( \text{op} ) is ( \text{PROJECT} )</td>
</tr>
<tr>
<td>15.</td>
<td>( s^i \leftarrow \text{SubstituteRef}(s^i, \text{ansrs}[\text{refs}[0]]) )</td>
</tr>
<tr>
<td>16.</td>
<td>( \text{ansrs}[i] \leftarrow \text{ans} )</td>
</tr>
<tr>
<td>17.</td>
<td>( \text{ansrs}[i] \leftarrow \text{ans} )</td>
</tr>
<tr>
<td>18.</td>
<td>return ( \text{ansrs}[\text{n}] )</td>
</tr>
</tbody>
</table>
extracted operator. For SELECT steps, we simply run the Answer(·) function. For PROJECT steps, we substitute the reference to the previous step in $s^i$ with its already computed answer, and then run Answer(·). For FILTER steps, we use a simple rule to extract a ‘normalized question’, $s^i$ from $s^i$ and get an intermediate answer $ans_{tmp}$ with Answer($s^i$). We then ‘intersect’ $ans_{tmp}$ with the referenced answer by multiplying the probabilities provided by the RC model and normalizing. For COMPARISON steps, we compare, with a discrete operation, the numbers returned by the referenced steps. The final answer is the highest probability answer of step $s^i$.

As our IR model we use bigram TF-IDF, proposed by Chen et al. (2017). Because the RC model is run on single-hop questions, we use the BERT-based RC model from Min et al. (2019b), trained solely on SQuAD (Rajpurkar et al., 2016).

BERTQA Baseline As BreakRC exploits question decompositions, we compare it with a model that does not. BERTQA receives as input the original natural language question, $x$. It uses the same IR model as BreakRC to retrieve paragraphs for $x$. For a fair comparison, we set its number of retrieved paragraphs such that it is identical to BreakRC (namely, 10 paragraphs for each QDMR step that involves IR). Similar to BreakRC, retrieved paragraphs are fed to a pretrained BERT-based RC model (Min et al., 2019b) to answer $x$. In contrast to BreakRC, that is trained on SQuAD, BERTQA is trained on the target dataset (HotpotQA), giving it an advantage over BreakRC.

A Combined Approach Last, we present an approach that combines the strengths of BreakRC and BERTQA. In this approach, we use the QDMR decomposition to improve retrieval only. Given a question $x$ and its QDMR $s$, we run BreakRC on $s$, but in addition to storing answers, we also store all the paragraphs retrieved by the IR model. We then run BERTQA on the question $x$ and the top-10 paragraphs retrieved by BreakRC, sorted by their IR ranking. This approach resembles that of Qi et al. (2019).

The advantage of Combined is that we do not need to develop an answering procedure for each QDMR operator separately, which involves different discrete operations such as comparison and intersection. Instead, we use BreakRC to retrieve contexts, and an end-to-end approach to learn how to answer the question directly. This can often handle operators not implemented in BreakRC, like BOOLEAN and UNION.

Table 6 shows model performance on HotpotQA. We report EM and $F_1$ using the official HotpotQA evaluation script. IR measures the percentage of examples in which the IR model successfully retrieved both of the ‘gold paragraphs’ necessary for answering the multi-hop question. To assess the potential utility of QDMR, we report results for BreakRC$^G$, which uses gold QDMRs, and BreakRC$^B$, which uses QDMRs predicted by a Copynet parser (§7.2).

Retrieving paragraphs with decomposed questions substantially improves the IR metric from 46.3 to 59.2 (BreakRC$^G$), or 52.5 (BreakRC$^B$). This leads to substantial gains in EM and $F_1$ for Combined$^G$ (43.3 to 52.4) and Combined$^B$ (43.3 to 49.3). The EM and $F_1$ of BreakRC$^G$ are only slightly higher than BERTQA because BreakRC does not handle certain operators, such as BOOLEAN steps (9.4% of the examples).

The majority of questions in HotpotQA combine select operations with either PROJECT (also called ‘bridge’ questions), COMPARISON, or FILTER. PROJECT and COMPARISON questions (Figure 6) were shown to be less susceptible

<table>
<thead>
<tr>
<th>Model</th>
<th>HotpotQA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EM</td>
</tr>
<tr>
<td>BERTQA</td>
<td>33.6</td>
</tr>
<tr>
<td>BreakRC$^B$</td>
<td>28.8</td>
</tr>
<tr>
<td>BreakRC$^G$</td>
<td>34.6</td>
</tr>
<tr>
<td>Combined$^B$</td>
<td>38.3</td>
</tr>
<tr>
<td>Combined$^G$</td>
<td>41.2</td>
</tr>
<tr>
<td>IR-NP</td>
<td>31.7</td>
</tr>
<tr>
<td>BreakRC$^B$</td>
<td>18.9</td>
</tr>
<tr>
<td>Combined$^B$</td>
<td>32.7</td>
</tr>
</tbody>
</table>

Table 6: Open-domain QA results on HotpotQA.
Figure 6: Examples of PROJECT and COMPARISON questions in HotpotQA (high-level QDMR).

<table>
<thead>
<tr>
<th>Model</th>
<th>PROJECT</th>
<th>COMPARISON</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EM  F₁</td>
<td>IR</td>
</tr>
<tr>
<td>BERTQA</td>
<td>22.8</td>
<td>31.0</td>
</tr>
<tr>
<td>BREAKRC⁺</td>
<td>25.4</td>
<td>33.7</td>
</tr>
<tr>
<td>BREAKRC⁻</td>
<td>32.2</td>
<td>41.9</td>
</tr>
</tbody>
</table>

Table 7: Results on PROJECT and COMPARISON questions from HotpotQA development set.

As QDMR structures can be easily annotated at scale, a natural question is how far are they from fully executable queries (known to be expensive to annotate). As shown in §4.3, QDMRs can be mapped to pseudo-logical forms with high precision (93.1%) by extracting formal operators and arguments from them. The pseudo-logical form differs from an executable query in the lack of grounding of its arguments (entities and relations) in KB constants. This stems from the design of QDMR as a domain-agnostic meaning representation (§2). QDMR abstracts away from a concrete KB schema by assuming an underlying ‘idealized’ KB, which contains all of its arguments.

Thus, QDMR can be viewed as an intermediate representation between a natural language question and an executable query. Such intermediate representations have already been discussed in prior work on semantic parsing. Kwiatkowski et al. (2013) and Choi et al. (2015) used underspecified logical forms as an intermediate representation. Guo et al. (2019) proposed a two-stage approach, separating between learning an intermediate text-to-SQL representation and the actual mapping to schema items. Works in the database community have particularly targeted the mapping of intermediate query representations into DB grounded queries, using schema mapping and join path inference (Androutsopoulos et al., 1995; Li et al., 2014; Baik et al., 2019). We argue that QDMR can be used as an easy-to-annotate representation in such semantic parsers, bridging between natural language and full logical forms.

7 QDMR Parsing

We now present evaluation metrics and models for mapping questions into QDMR structures.

Task Definition Given a question \(x\) we wish to map it to its QDMR steps, \(s = (s^1, \ldots, s^n)\). One can frame this as a sequence-to-sequence problem where \(x\) is mapped to a string representing its decomposition. We add a special separating token (SEP), and define the target string to be \(s^1, \ldots, s_{m_1}, \text{(SEP)}, s^2, \ldots, s_{m_2}, \text{(SEP)}, \ldots, s^n_{m_n}\), where \(m_1, \ldots, m_n\) are the number of tokens in each decomposition step.

7 Issuing an IR query over each "content word" in the question, instead of each noun phrase, led to poor results.
7.1 Evaluation Metrics

We wish to assess the quality of a predicted QDMR, \( \hat{s} \) to a gold standard, \( s \). Figure 7 lists various properties by which question decompositions may differ, such as granularity (e.g., steps 1–3 of decomposition 1 are merged into the first step of decomposition 2), ordering (e.g., the last two steps are swapped) and wording (e.g., using ‘‘from’’ instead of ‘‘on’’). While such differences do not affect the overall semantics, the second decomposition can be further decomposed. To measure such variations, we introduce two types of evaluation metrics. Sequence-based metrics treat the decomposition as a sequence of tokens, applying standard text generation metrics. As such metrics ignore the QDMR graph structure, we also use graph-based metrics that compare the predicted graph \( G_{\hat{s}} \) to the gold QDMR graph \( G_s \) (see §2).

**Sequence-based scores**, where higher values are better, are denoted by \( \uparrow \). Graph-based scores, where lower values are better, are denoted by \( \downarrow \).

- **Exact Match** \( \uparrow \): Measures exact match between \( s \) and \( \hat{s} \), either 0 or 1.
- **SARI** \( \uparrow \) (Xu et al., 2016): SARI is commonly used in tasks such as text simplification. Given \( s \), we consider the sets of added, deleted, and kept n-grams when mapping the question \( x \) to \( s \). We compute these three sets for both \( s \) and \( \hat{s} \) using the standard of up to 4-grams, then average (a) the F\(_1\) for added n-grams between \( s \) and \( \hat{s} \), (b) the F\(_1\) for kept n-grams, and (c) the precision for the deleted n-grams.
- **Graph Edit Distance (GED)** \( \downarrow \): A graph edit path is a sequence of node and edge edit operations (addition, deletion, and substitution), where each operation has a predefined cost. GED computes the minimal-cost graph edit path required for transitioning from \( G_s \) to \( G_{\hat{s}} \) (and vice versa), normalized by \( \max(\{|G_s|, |G_{\hat{s}}|\}) \). Operation costs are 1 for insertion and deletion of nodes and edges. The substitution cost of two nodes \( u, v \) is set to be \( 1 - \text{Align}(u, v) \), where \( \text{Align}(u, v) \) is the ratio of aligned tokens between these steps.
- **GED+** \( \downarrow \): Comparing the QDMR graphs in Figure 8, we consider the splitting and merging of graph nodes. We implement GED+, a variant of GED with additional operations to merge (split) a set of nodes (node), based on the A* algorithm (Hart et al., 1968).\(^8\)

7.2 QDMR Parsing Models

We present models for QDMR parsing, built over AllenNLP (Gardner et al., 2017).

- **COPY**: A model that copies the input question \( x \), without introducing any modifications.
- **RULEBASED**: We defined 12 decomposition rules, to be applied over the dependency tree of the question, augmented with coreference relations. A rule is a regular expression over the question dependency tree, which invokes a decomposition operation when matched (Table 8). For example, the rule for relative clauses (\( \text{relcl} \)) breaks the question at the relative pronoun ‘‘that’’, while adding a reference to the preceding part of the sentence. A full decomposition is obtained by recursively applying the rules until no rule is matched.
- **SEQ2SEQ**: A sequence-to-sequence neural model with a 5-layer LSTM encoder and attention at decoding time.

\(^8\)Because of its exponential worst-case complexity, we compute GED+ only for graphs with up to 5 nodes, covering 75.2% of the examples in the development set of Break.
Table 8: The decomposition rules of RuleBased. Rules are based on dependency labels, part-of-speech tags and coreference edges. Text fragments used for decomposition are in boldface.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>be-auxpass</td>
<td>How many objects smaller than the matte object are silver [objects smaller than the matte object. How many #1 silver]</td>
</tr>
<tr>
<td>do-subj</td>
<td>Find the average rating star for each movie that are not reviewed by Brittany Harris. [Brittany Harris, the average rating star for each movie that not reviewed by #1]</td>
</tr>
<tr>
<td>sub-verb</td>
<td>Year did Tim Howard win the Superbowl? [Tim Howard win the Superbowl]</td>
</tr>
<tr>
<td>conjunction</td>
<td>Who trades with China and has a capital city called Khartoum? [Who has a capital city called Khartoum. #1 trades with China]</td>
</tr>
<tr>
<td>how-many</td>
<td>How many metallic objects appear in this image? [metallic objects appear in this image, the number of #1]</td>
</tr>
<tr>
<td>single-prep</td>
<td>Find the rail of the problems reported after 1978. [the problems reported after 1978, ids of #1]</td>
</tr>
<tr>
<td>multi-prep</td>
<td>Find objects from Tacoma to Orlando. #1 on Saturday [Rails, #1 from Tacoma. #2 to Orlando. #3 on Saturday]</td>
</tr>
<tr>
<td>rel</td>
<td>Find all the students that do not have a back vocal [all the songs, #1 that do not have a back vocal]</td>
</tr>
<tr>
<td>superlative</td>
<td>What is the smallest state bordering Ohio? [state bordering ohio, the smallest #1]</td>
</tr>
<tr>
<td>acl</td>
<td>what flights from Tacoma to Orlando, on Saturday [flights, #1 from Tacoma. #2 to Orlando. #3 on Saturday]</td>
</tr>
<tr>
<td>sent-coref</td>
<td>The claim that has the largest total settlement amount. Return the effective date of the claim. [the claim that has the largest total settlement amount, the effective date of #1]</td>
</tr>
</tbody>
</table>

For high-level examples (from RC datasets), as questions are often less structured, they require a deeper semantic understating from the decomposition model. Only 8% of the predictions were an exact match, with an additional 46% being semantically equivalent to the gold. The remaining 46% were of erroneous predictions (see Table 10).

8 Related Work

Question Decomposition Recent work on QA through question decomposition has focused mostly on single modalities (Gupta and Lewis, 2018; Guo et al., 2019; Min et al., 2019b). QA using neural modular networks has been suggested for both KBs and images by Andreas et al. (2016) and Hu et al. (2017). Question decomposition over text was proposed by Talmor and Berant (2018), however over a much more limited set of questions than in Break. Iyyer et al. (2017) have also decomposed questions to create a “sequential question answering” task. Their annotators viewed a web table and performed actions over it to retrieve the cells that constituted the answer. Conversely, we provided annotators only with the question, as QDMR is agnostic to the original context.

An opposite annotation cycle to ours was presented in Cheng et al. (2018). The authors generate sequences of simple questions which crowdlab participants paraphrase into a compositional question. Questions in Break are composed by humans, and are then decomposed to QDMR.

Semantic Formalism Annotation Labeling corpora with a semantic formalism has often been reserved for expert annotators (Dahl et al., 1994; Zelle and Mooney, 1996; Abend and Rappoport, 2013; Yu et al., 2018). Recent work has focused on cheaply eliciting quality annotations from non-experts through crowdsourcing (He et al., 2016; Iyer et al., 2017; Michael et al., 2018). FitzGerald et al. (2018) facilitated non-expert annotation by introducing a formalism expressed in natural language for semantic-role-labeling. This mirrors QDMR, as both are expressed in natural language.

Relation to Other Formalisms QDMR is related to Dependency-based Compositional Semantics (Liang et al., 2013), as both focus on question representations. However, QDMR is designed
Table 9: Performance of QDMR parsing models on the development and test set. GED+ is computed only for the subset of QDMR graphs with up to 5 nodes, covering 66.1% of QDMRs and 97.6% of high-level data.

<table>
<thead>
<tr>
<th>Question</th>
<th>Gold</th>
<th>Prediction (COPYNET)</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>How many objects other than cylinders are there?</td>
<td>(1) objects; (2) cylinders; (3) #1 besides #2; (4) number of #3.</td>
<td>(1) objects; (2) cylinders; (3) number of #1; (4) number of #2; (5) difference of #3 and #4.</td>
<td>sem. equiv. (30%)</td>
</tr>
<tr>
<td>Where is the youngest teacher from?</td>
<td>(1) teachers; (2) the youngest of #1; (3) where is #2 from.</td>
<td>(1) youngest teacher; (2) where is #1.</td>
<td>incorrect</td>
</tr>
<tr>
<td>Kyle York is the Chief Strategy Officer of a company acquired by what corporation in 2016?</td>
<td>(1) company that Kyle York is the Chief Strategy Officer of; (2) corporation that acquired #1 in 2016.</td>
<td>(1) company that Kyle York is the Chief Strategy Officer of; (2) corporation in 2016 that #1 was acquired by.</td>
<td>sem. equiv. (46%)</td>
</tr>
<tr>
<td>Dayton’s Devils had a cameo from the ‘MASH’ star who played what role on the show?</td>
<td>(1) MASH star that Dayton’s Devils had a cameo from; (2) role that #1 played on the show.</td>
<td>(1) the MASH that Dayton’s Devils had a cameo; (2) what role on the show star of #1 played.</td>
<td>incorrect</td>
</tr>
</tbody>
</table>

Table 10: Manual error analysis of the COPYNET model predictions. Lower examples are of high-level QDMRs.

## 9 Conclusion

In this paper, we presented a formalism for question understanding. We have shown it is possible to train crowd-workers to produce such representations with high quality at scale, and created BREAK, a benchmark for question decomposition with over 83K decompositions of questions from 10 datasets and 3 modalities (DB, images, text). We presented the utility of QDMR for both open-domain question answering and semantic parsing, and constructed a QDMR parser with reasonable performance. QDMR proposes a promising direction for modeling question understanding, which we believe will be useful for multiple tasks in which reasoning is probed through questions.

### Acknowledgments

This work was completed in partial fulfillment for the PhD of Tomer Wolfson. This research was partially supported by The Israel Science Foundation (grants 942/16 and 978/17), and The Yandex Initiative for Machine Learning and the European Research Council (ERC) under the European Union Horizons 2020 research and innovation programme (grant ERC DELPHI 802800).

### References


Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In 7th Linguistic Annotation Workshop and Interoperability with Discourse.


James Clarke, Dan Goldwasser, Ming-Wei Chang, and Dan Roth. 2010. Driving semantic parsing from the world’s response. In Computational Natural Language Learning (CoNLL), pages 18–27.


Wen-tau Yih, Matthew Richardson, Christopher Meek, Ming-Wei Chang, and Jina Suh. 2016. The value of semantic parse labeling for knowledge base question answering. In Association for Computational Linguistics (ACL).

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir R. Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In Empirical Methods in Natural Language Processing (EMNLP).
