Summary: We present comprehensive analysis on information-seeking QA datasets, namely Natural Questions [1] and TyDi QA [2].

- Our **Gold-Type & Gold-Paragraph** experiments show answerability prediction and paragraph retrieval remain hard.
- Our **Answerability Prediction** experiments shows unanswerable questions in information-seeking QA have unique characteristics, which make answerability prediction challenging even given gold context.
- Our **Unanswerability Annotation** reveal where the unanswerability arise and how we could improve answer coverage.

Background

- **Machine Reading Comprehension** questions are written by annotators who know the answers given a paragraph.
- **Information-seeking QA** questions are written by annotators who don’t know the answers.

Answerability Prediction Experiments

**Task setup**: predicting if a question is answerable or not

- **Question-only**: only use question input
- **Full-input**: use both a question and a document

**Datasets**
- Natural Questions, TyDi QA
- SQuAD 2.0 (MRC)

**Models**
- **Majority**
- **Question-only**
- **QA models**: ETC for NQ, RetroReader [9] for SQuAD 2.0, multilingual BERT-based model for TyDi.
- **Human**

**Results**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Majority</th>
<th>Question Only</th>
<th>QA Model</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td>58.9</td>
<td>72.7</td>
<td>82.5</td>
<td>85.6</td>
</tr>
<tr>
<td>TyDi QA</td>
<td>58.2</td>
<td>70.2</td>
<td>79.4</td>
<td>94.0</td>
</tr>
<tr>
<td>SQuAD 2.0</td>
<td>50.0</td>
<td>63.0</td>
<td>84.1</td>
<td></td>
</tr>
</tbody>
</table>

Unanswerable questions both Q-only & QA models detect

- obviously too vague or invalid
- key phrase
- complex reasoning

Gold Type / Gold Paragraph Experiments

**Settings**

- **Gold Type**: provides the answer type (answerability prediction)
- **Gold Paragraph**: provides the paragraph (paragraph retrieval)
- **Gold Type + Gold Paragraph**

**Models**

- **ETC** [7] for NQ.
- **Multilingual BERT-based model** [5, 8] for TyDi QA

Human performance

- **Super** (25 way), **Single** (5 way) for NQ
- **3-way** human annotation for TyDi QA

Discussion: How we can improve answer coverage?

1. **New Task formulation**: extractive QA over single document may not cover all of the diverse information-seeking questions.
   - Generative answer, multi-paragraph evidence
2. **Alternative knowledge source**: improve answer coverage by using additional knowledge sources.

Annotation Unanswerability – Category & Statistics

- **Factoid question (retriever failure)**
- **Non-factoid question (formula failure)**
- **Multi-evidence question (formula failure)**
- **Invalid question (bad question)**
- **False premise (bad question)**
- **Invalid answers (annotation error)**

Gold Type / Gold Paragraph Experiments

- **Results on NQ**
  - Long Answer F1: 90% (Gold Type) vs. 88% (Gold Paragraph & Paragraph)
  - Short Answer F1: 95% (Gold Type) vs. 90% (Gold Paragraph & Paragraph)

- **Results on TyDi QA**
  - Long Answer F1: 65% (Gold Type) vs. 60% (Gold Paragraph & Paragraph)
  - Short Answer F1: 75% (Gold Type) vs. 70% (Gold Paragraph & Paragraph)

Table / Infobox

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Gold Type</th>
<th>Gold Paragraph &amp; Paragraph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td>58.9</td>
<td>72.7</td>
</tr>
<tr>
<td>TyDi QA</td>
<td>58.2</td>
<td>70.2</td>
</tr>
<tr>
<td>SQuAD 2.0</td>
<td>50.0</td>
<td>63.0</td>
</tr>
</tbody>
</table>