Challenges in Information-seeking QA:
Unanswerable Questions and Paragraph Retrieval

Akari Asai
University of Washington

Eunsol Choi
The University of Texas at Austin

Outlines

- **Background**
- **Gold Type / Gold Paragraph** experiments

  What are the remaining head-rooms in information-seeking QA?
- **Answerability Prediction** experiments

  How well do our models perform on answerability prediction?
- **Unanswerability Annotation**

  What makes the questions unanswerable?
- **Discussions**

  How can we improve answer coverage?

Background: Machine reading Comprehension (MRC)

- **Machine Reading Comprehension** questions are written by annotators who know the answers to test the models' ability to comprehend a single paragraph.

Question (Q) + Paragraph (P) → Short Answer (A)

What was Maria Curie the first female recipient of?

One of the most famous people born in Warsaw was Maria Skłodowska-Curie, who was the first female recipient of the Nobel Prize.

→ Nobel Prize

Overlap between P and Q

Supporting evidence

Background: Information-seeking QA

- **Information-seeking QA** are the questions that are written by annotators who do not know the answers independently from existing documents.

Question (Q) + Document (D) → Short Answer (A)

Long answer / Gold paragraph (GP)

who is the voice of tony the tiger?

Thurl Arthur Ravenscroft

Thurl Ravenscroft - en.wikipedia

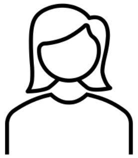
Thurl Arthur Ravenscroft was an American actor and bass singer known as the booming voice behind Kellogg's Frosted Flakes animated spokesman Tony the Tiger for more than five decades.

His voice acting career began in 1940 and lasted until his death in 2005 at age 91.

Background: Natural Questions & TyDi QA

- **Natural Questions** are the collections of anonymized Google Search queries.
- **TyDi QA** asks native speakers to author questions they are interested in.

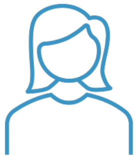
NQ



who is the voice of tony the tiger?

Google Search User

TyDi

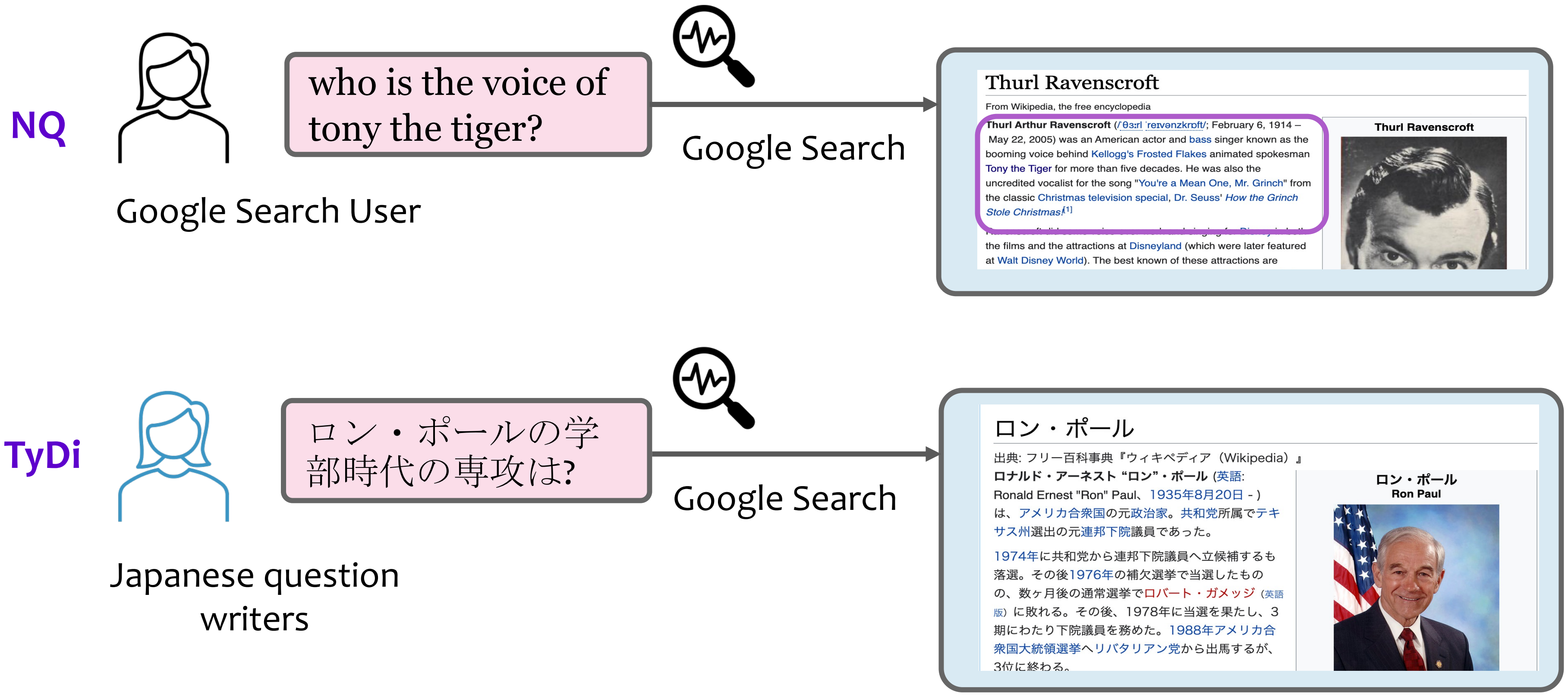


ロン・ポールの学部時代の専攻は?

Japanese question writers

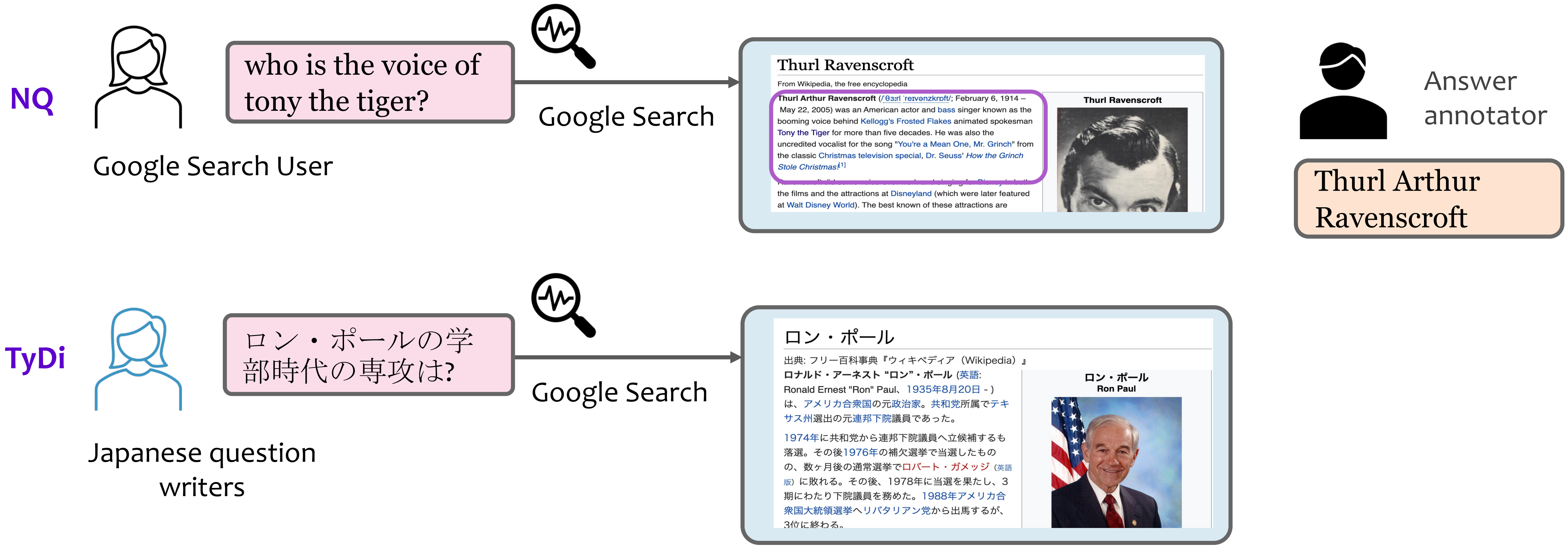
Background: Natural Questions & TyDi QA

- **Natural Questions** are the collections of anonymized Google Search queries.
- **TyDi QA** asks native speakers to author questions they are interested in.
- Google Search Top 1 Wikipedia articles are used as Document.



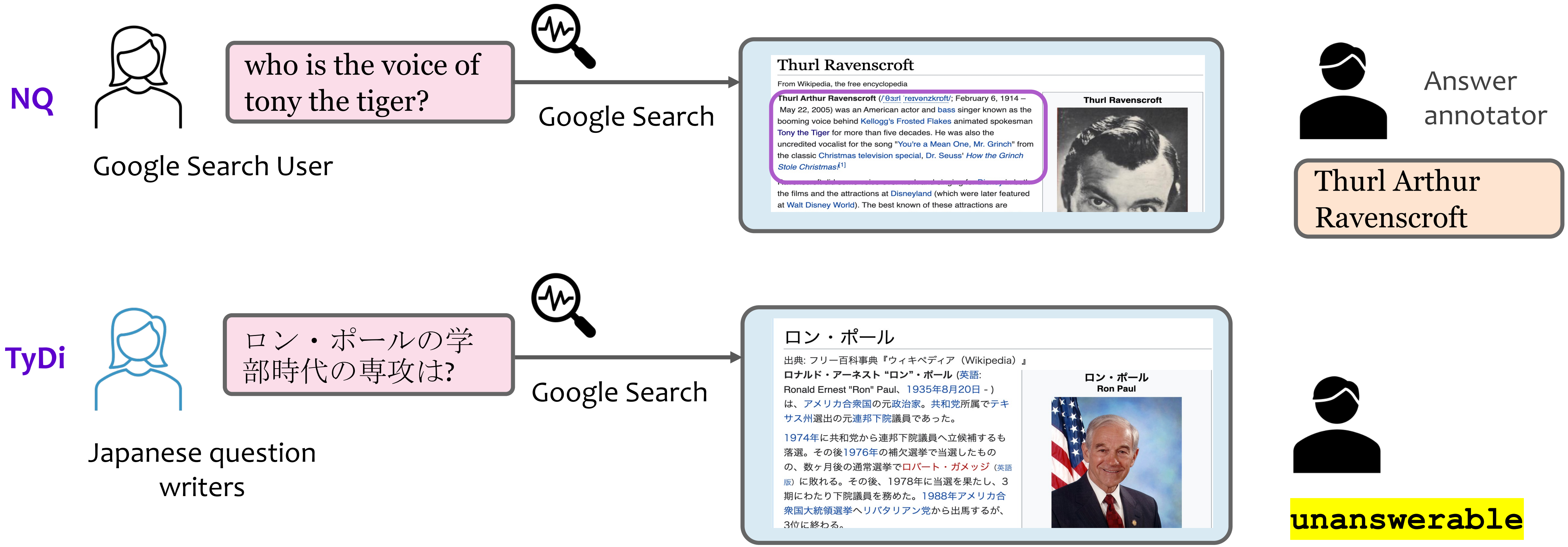
Background: Natural Questions & TyDi QA

- **Natural Questions** are the collections of anonymized Google Search queries.
- **TyDi QA** asks native speakers to author questions they are interested in.
- Google Search Top 1 Wikipedia articles are used as Document.



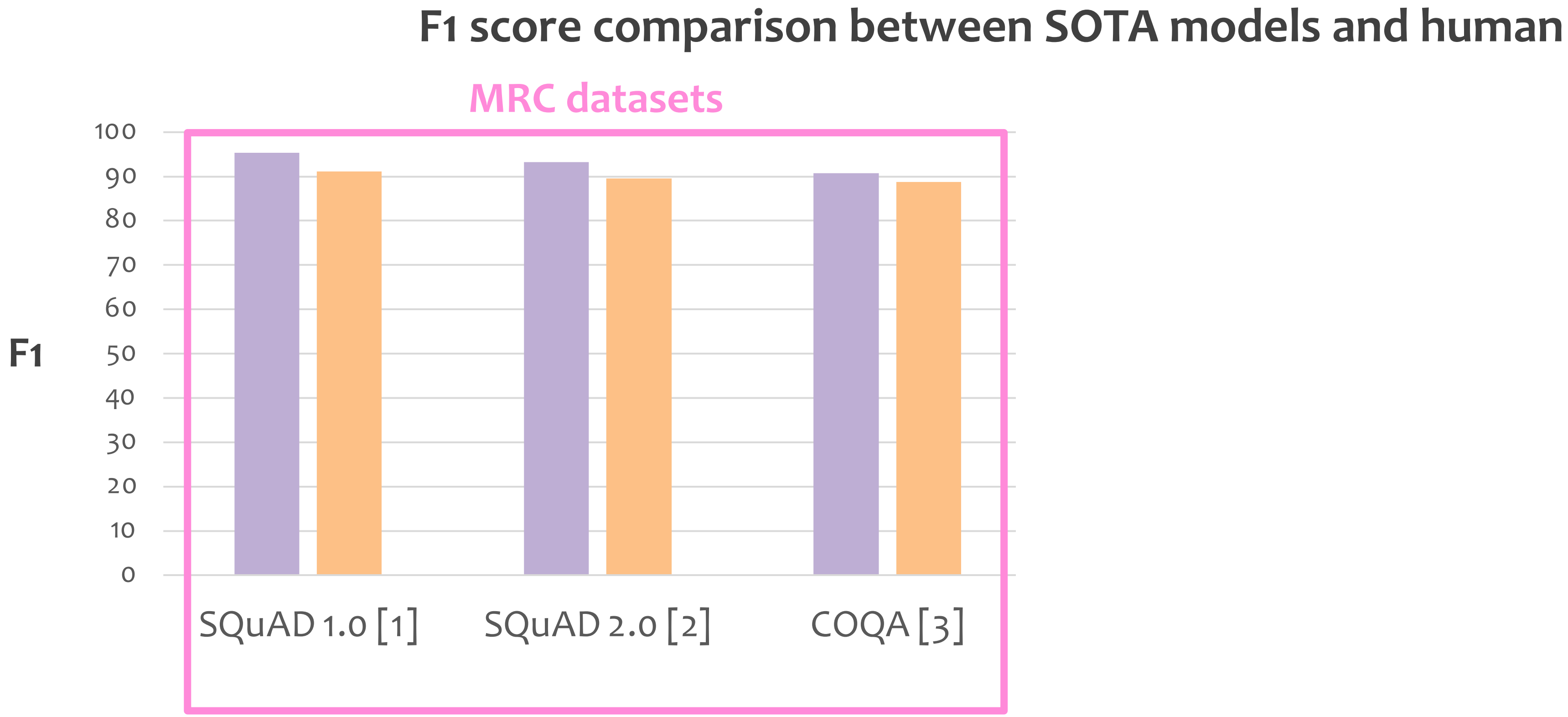
Background: Natural Questions & TyDi QA

- **Natural Questions** are the collections of anonymized Google Search queries.
- **TyDi QA** asks native speakers to author questions they are interested in.
- Google Search Top 1 Wikipedia articles are used as Document.



Background: Our models are better than human?

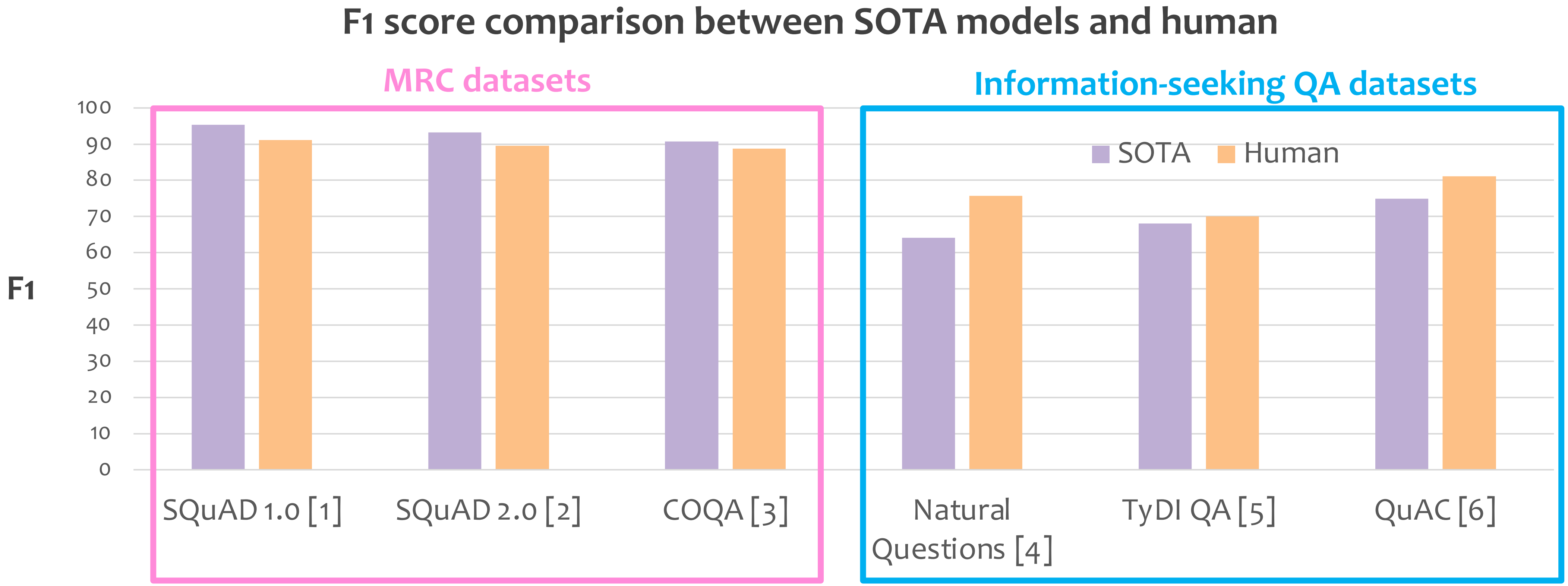
- In MRC datasets, the state-of-the-art models outperform human performance.



[1] Rajpurkar et al. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *EMNLP*.
[2] Rajpurkar et al. (2018). Know What You Don't Know: Unanswerable questions for SQuAD. In *ACL*.
[3] Reddy et al. (2019) CoQA: Conversational Question Answering Challenge . *TACL*.

Background: Our models are better than human?

- In MRC datasets, the state-of-the-art models outperform human performance.
- State-of-the-art models still struggles in Information-seeking QA.



[1] Rajpurkar et al. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *EMNLP*.
[2] Rajpurkar et al. (2018). Know What You Don't Know: Unanswerable questions for SQuAD. In *ACL*.
[3] Reddy et al. (2019) CoQA: Conversational Question Answering Challenge . *TACL*.
[4] Kwiatkowski et al. (2019). Natural Questions: a Benchmark for Question Answering. *TACL*.
[5] Clark et al. (2020): TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. TACL.
[6]Choi et al. (2018). QuAC: Question Answering in Context. In *EMNLP*.

Background: Information-seeking QA v.s. MRC

- **Answerability perdition** and **paragraph retrieval** is required.
- SQuAD 2.0 has **unanswerable questions** but they are written with evidence context provided unlike TyDi QA / Natural Questions.

	Information seeking QA		MRC	
	Natural Questions	TyDi QA	SQuAD	SQuAD 2.0
Limited Q-P lexical overlap	✓	✓		
% of Unanswerable Questions	**50.1**	**59.9**	0	**33.4** (artificially created)
Avg. number of paragraphs	131.3	41.1	1	1

Contributions

We present the first comprehensive analysis on information-seeking QA datasets, namely Natural Questions and TyDi QA.

- Our controlled experiments show **Answerability Prediction** and **Paragraph retrieval remain challenging.**

- Our **unanswerability annotations** revel the unique cause of unanswerability and how we can improve future dataset development and task formulation.

Outlines

- **Background**
- **Gold Type** / **Gold Paragraph** experiments

  What are the remaining head-rooms in information-seeking QA?
- **Answerability Prediction** experiments

  How well do our model performs on answerability prediction?
- **Unanswerability Annotation**

  What makes the questions unanswerable?
- **Discussions**

  How can we improve answer coverage?

Gold-Type, Paragraph: Remaining head-rooms

- We conduct two controlled experiments
  - **Gold Type**: provide the type, `short` / `long` / `unanswerable`

Gold Type

who is the voice of tony the tiger?

+

Thurl Arthur Ravenscroft was an American actor known as Kellogg's Frosted Flakes animated spokesman Tony the Tiger .

His voice acting career began in 1940.

+

`short`

Gold-Type, Paragraph: Remaining head-rooms

- We conduct two controlled experiments
  - **Gold Type**: provide the type, `short` / `long` / `unanswerable`
  - **Gold Paragraph**: provide the gold paragraph

Gold Type

who is the voice of tony the tiger?

+

Thurl Arthur Ravenscroft was an American actor known as Kellogg's Frosted Flakes animated spokesman Tony the Tiger .

His voice acting career began in 1940.

+

`short`

Gold Paragraph

who is the voice of tony the tiger?

+

Thurl Arthur Ravenscroft was an American actor known as Kellogg's Frosted Flakes animated spokesman Tony the Tiger .

Gold-Type, Paragraph: Remaining head-rooms

- We conduct two controlled experiments
  - **Gold Type + Gold Paragraph**: provide both the gold paragraph and the gold type

Gold Type + Gold Paragraph

who is the voice of tony the tiger?

+

Thurl Arthur Ravenscroft was an American actor known as Kellogg's Frosted Flakes animated spokesman Tony the Tiger .

+

`short`

Gold-Type, Paragraph: Models

- Models – Use state-of-the-art models in NQ and TyDi QA
  - **ETC [7]**: can take up to 4k tokens, SOTA for NQ
  - **Multilingual BERT-based baselines [5, 8]**: predict no-answer scores, TyDi QA's best baseline.

[7] Ainslie et al. (2020). ETC: Encoding Long and Structured Inputs in Transformers. In *EMNLP*.
[8]Alberti et al. (2019). A BERT Baseline for the Natural Questions.

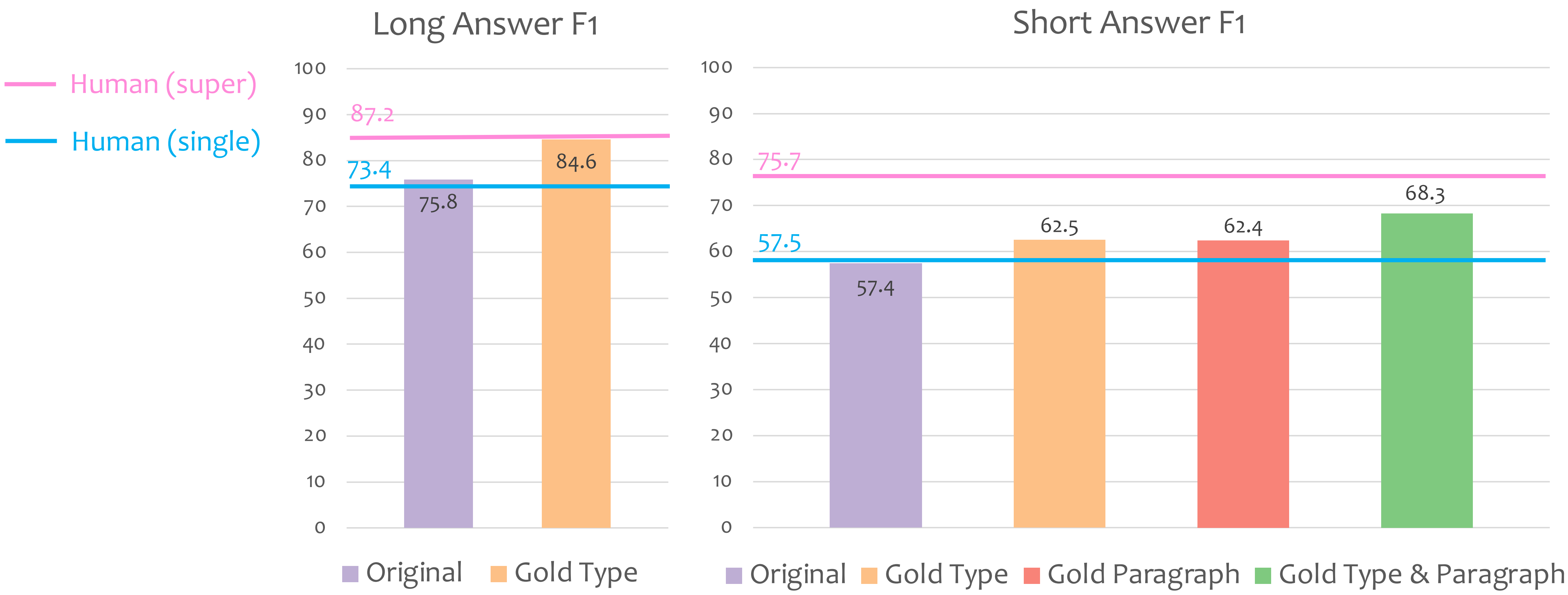
Gold-Type, Paragraph: Models

- Models – Use state-of-the-art models in NQ and TyDi QA
  - **ETC** [7]: can take up to 4k tokens, SOTA for NQ
  - **Multilingual BERT-based baselines** [5, 8]: predict no-answer scores, TyDi QA's best baseline.

- Human performance
  - **Single** (NQ): performance of single annotator
  - **Super** (NQ): performance of collected 25 annotators (NQ's upper bound)
  - **Three-way** (TyDi): performance of three annotators

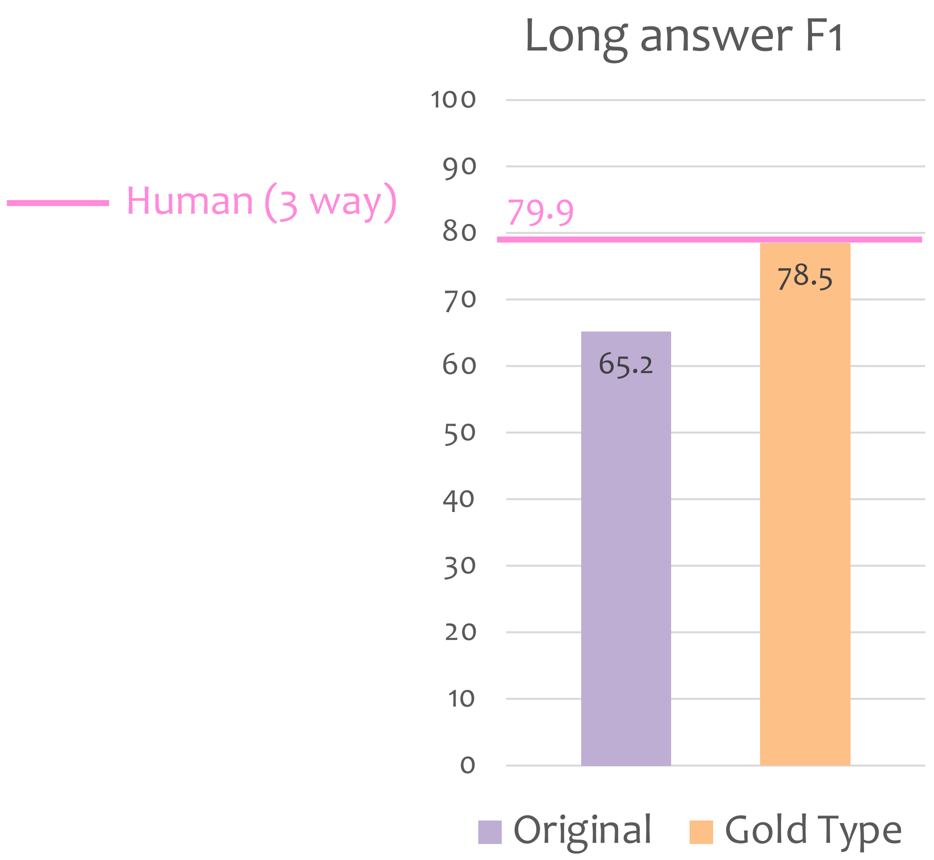
Gold-Type, Paragraph: Results on NQ

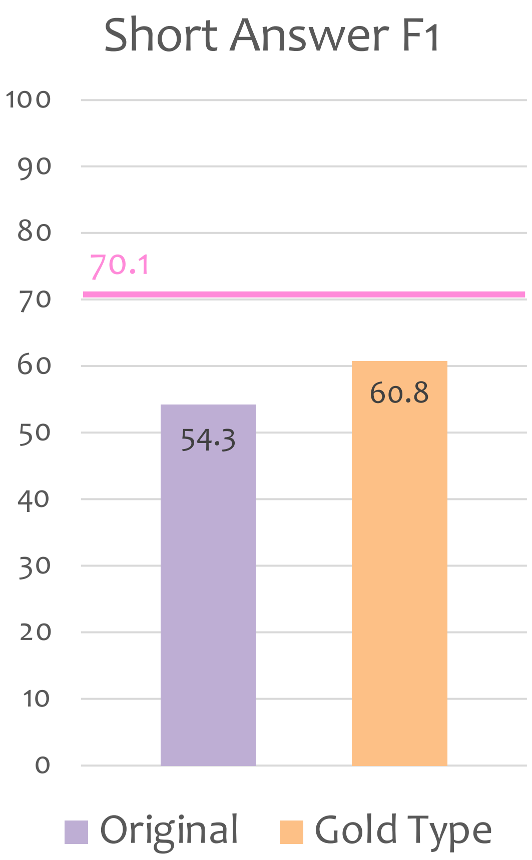
- Given gold type and gold paragraph information, the state-of-the-art models outperform single human performance on NQ.



Gold-Type, Paragraph: Results on TyDi QA

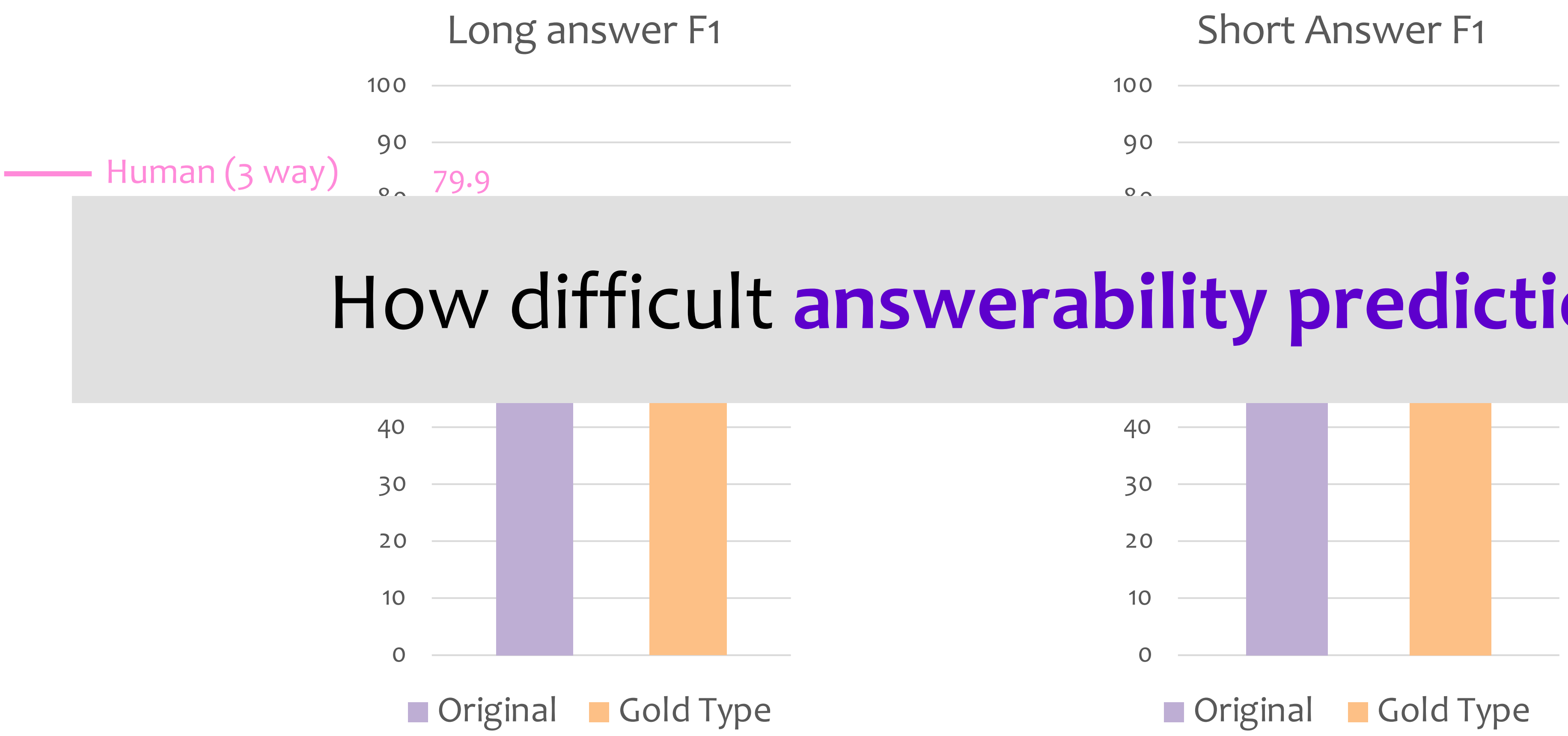
- Both Gold Type and Gold Paragraph contribute to performance improvements on TyDi QA as well.





Gold-Type, Paragraph: Results on TyDi QA

- Both Gold Type and Gold Paragraph contribute to performance improvements on TyDi QA as well.



How difficult **answerability prediction** is?

Outlines

- **Background**
- **Gold Type / Gold Paragraph** experiments

  What are the remaining head-rooms in information-seeking QA?
- **Answerability Prediction** experiments

  How well do our models perform on answerability prediction?
- **Unanswerability Annotation**

  What makes the questions unanswerable?
- **Discussions**

  How can we improve answer coverage?

Answerability Prediction: Task setup

- **Question-only setting**

who is the voice of tony the tiger?

Model

✓ Answerable

No answer

Answerability Prediction: Task setup

- **Question-only setting**

who is the voice of tony the tiger?

Model

✓ Answerable

No answer

- **Full input setting**

who is the voice of tony the tiger?

Thurl Arthur Ravenscroft was an American actor known as Kellogg's Frosted Flakes animated spokesman Tony the Tiger .

His voice acting career began in 1940.

Model

✓ Answerable

No answer

Answerability Prediction: Datasets and Models

- **Information-seeking QA:** Natural Questions & TyDi QA
- **Machine Reading Comprehension** : SQuAD 2.0
  - Questions are created by annotators who try to confuse models.

Answerability Prediction: Models

- Partial input
  - **Majority**: output the majority class in training data
  - **Question-Only:** two-way classification model based on BERT
- Full information
  - **SOTA QA models**
  - **Human**

Answerability Prediction: Comparison w/ SQuAD 2.0

- Question-only particularly struggles in SQuAD 2.0.

Dataset	Majority	Question Only	QA model	Human
Natural questions	58.9	72.7	82.5	85.6
TyDi QA	58.2	70.2	79.4	94.0
SQuAD 2.0	50.0	63.0	94.1	--

Answerability Prediction: Comparison w/ SQuAD 2.0

- Question-only particularly struggles in SQuAD 2.0.
- QA model shows much better results than in NQ & TyDi → State-of-the-art model already easily identify the unanswerable questions in SQuAD 2.0

Dataset	Majority	Question Only	QA model	Human
Natural questions	58.9	72.7	82.5	85.6
TyDi QA	58.2	70.2	79.4	94.0
SQuAD 2.0	50.0	63.0	94.1	--

Answerability Prediction: Comparison w/ SQuAD 2.0

- Many SQuAD 2.0 questions become unanswerable because of **incorrect entity names**, **false premise**, **context negation and missing information** [9].
- These can be easily solved by matching a question and a short paragraph.

False premise

When was Warsaw ranked as the 22nd most liveable city in the world?

In 2012 the Economist Intelligence Unit ranked Warsaw as the **32nd** most liveable city in the world.

Incorrect entity names (Entity salad)

How many people live in Carpathia?

Warsaw .. from the **Carpathian** Mountain. Warsaw's population is estimated at 1.740 million, which makes Warsaw the 9th most-populous capital city in the European Union

[9] Yatskar (2018). A Qualitative Comparison of CoQA, SQuAD 2.0 and QuAC. In *NAACL*.

Answerability Prediction: Comparison w/ SQuAD 2.0

- Many SQuAD 2.0 questions become unanswerable because of **incorrect entity names**, **false premise**, **context negation and missing information** [9].
- These can be easily solved by matching a question and a short paragraph.

False premise

When was Warsaw ranked as the 22nd most liveable city in the world?

In 2012 the Economist Intelligence Unit ranked Warsaw as the **32nd** most liveable city in the world.

Incorrect entity names (Entity salad)

How many people live in Carpathia?

Warsaw .. from the **Carpathian** Mountain. Warsaw's population is estimated at 1.740 million, which makes Warsaw the 9th most-populous capital city in the European Union

→ What makes information-seeking questions unanswerable?

[9] Yatskar (2018). A Qualitative Comparison of CoQA, SQuAD 2.0 and QuAC. In *NAACL*.

Outlines

- **Background**
- **Gold Type / Gold Paragraph** experiments

  What are the remaining head-rooms in information-seeking QA?
- **Answerability Prediction** experiments

  How well do our models perform on answerability prediction?
- **Unanswerability Annotation**

  What makes the questions unanswerable?
- **Discussions**

  How can we improve answer coverage?

Annotating Unanswerability: Categories

We define several categories of unanswerability in information-seeking QA.

[1] **Retrieval Miss**: paired with a document that do not contain a single gold paragraph and answer.

[2] **Invalid QA**: Annotated QA data is partially incorrect.

Annotating Unanswerability: Categories

We define several categories of unanswerability in information-seeking QA.

[1] **Retrieval Miss:** paired with a document that do not contain a single gold paragraph and answer.

Factoid question (retriever failure)

Non-Factoid question (formulation failure)

Multi-evidence question (formulation failure)

[2] **Invalid QA:** Annotated QA data is partially incorrect.

Annotating Unanswerability: Categories

We define several categories of unanswerability in information-seeking QA.

[1] **Retrieval Miss:** paired with a document that do not contain a single gold paragraph and answer.

Factoid question (retriever failure)

Non-Factoid question (formulation failure)

Multi-evidence question (formulation failure)

[2] **Invalid QA:** Annotated QA data is partially incorrect.

Invalid question (bad question)

False Premise (bad question)

Invalid answers (annotation error)

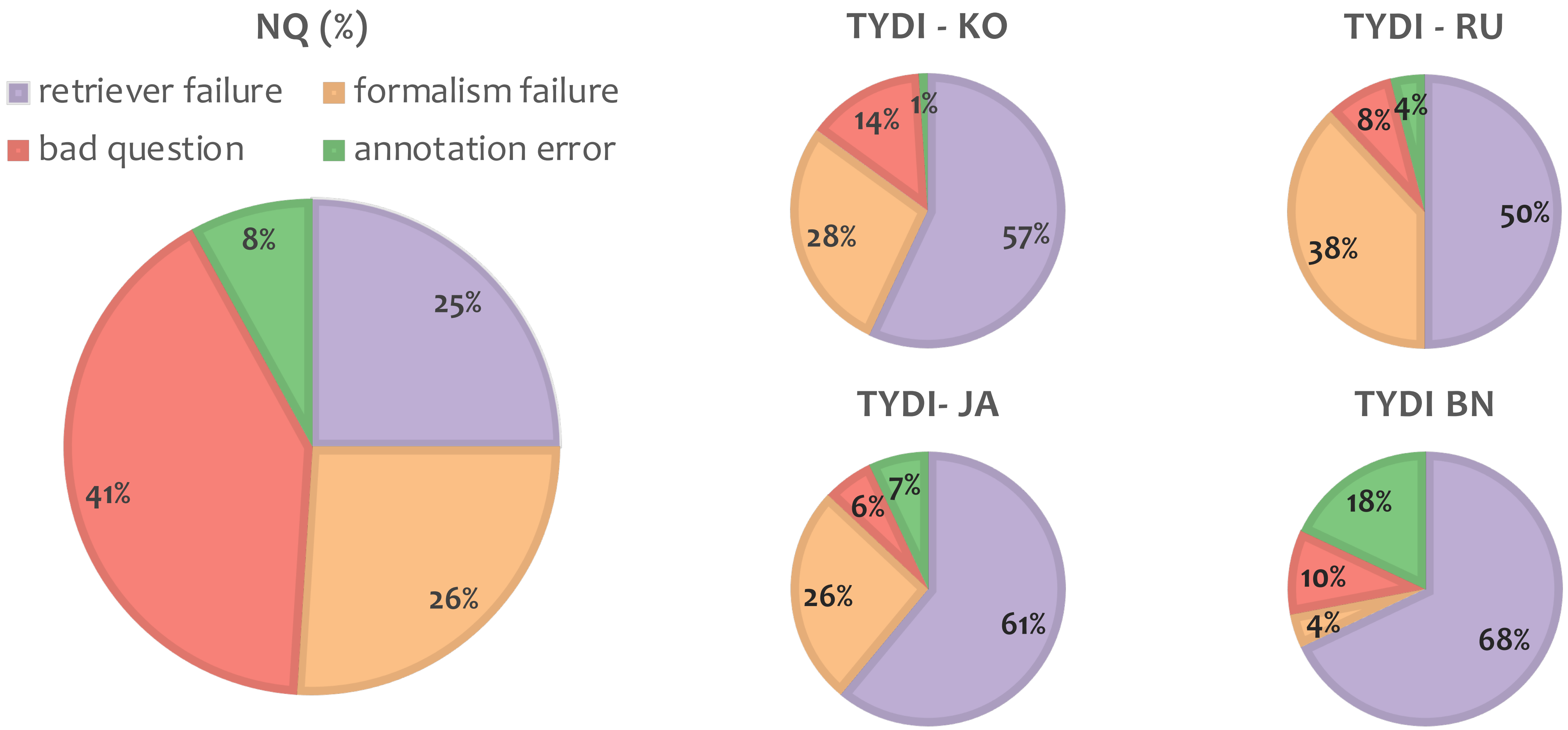
Annotating Unanswerability: Statistics

- We annotated 800 information-seeking questions from NQ and TyDi QA across six languages.

Dataset	Language	Numbers
Natural Questions	English	450
TyDi QA	Bengali	50
TyDi QA	Japanese	100
TyDi QA	Korean	100
TyDi QA	Russian	50
TyDi QA	Telugu	50

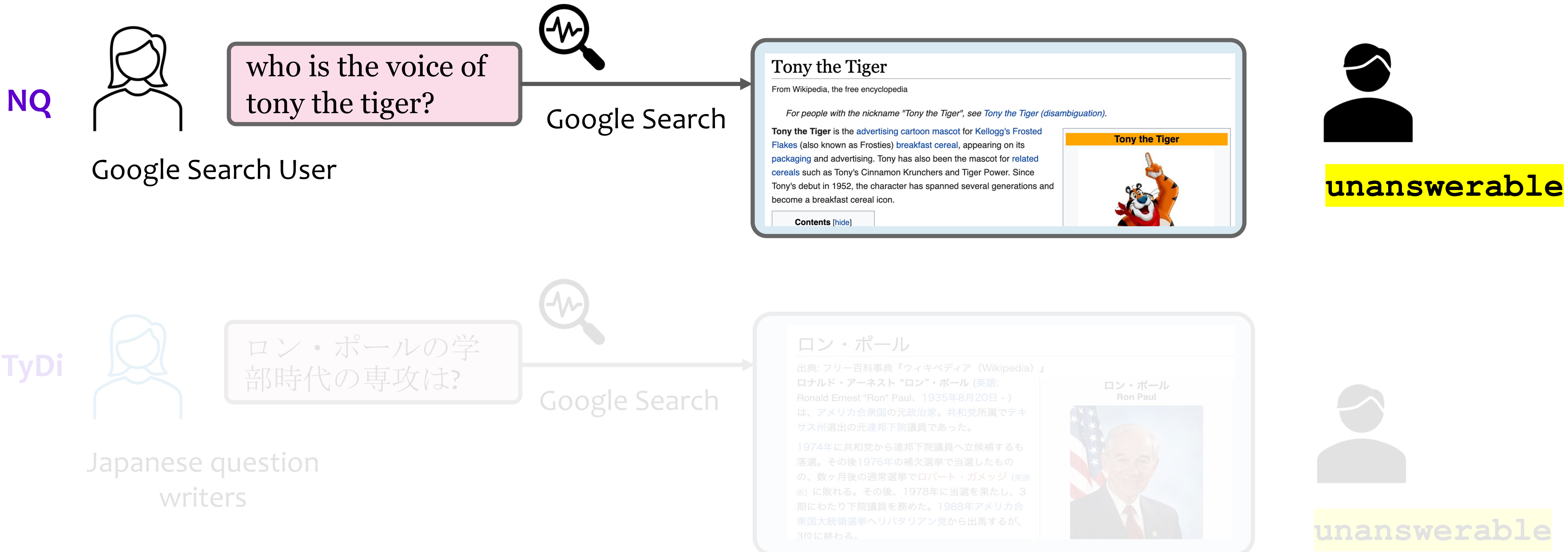
Annotating Unanswerability: Results

- More retrieval failures in TyDi QA.
- More bad questions in NQ due to ambiguous or ill-formed questions.



Annotating Unanswerability: Retriever failure

- Quality of the retriever (Search Engine): retrieved document is not most relevant
- Missing information: No article with enough information in the target language.

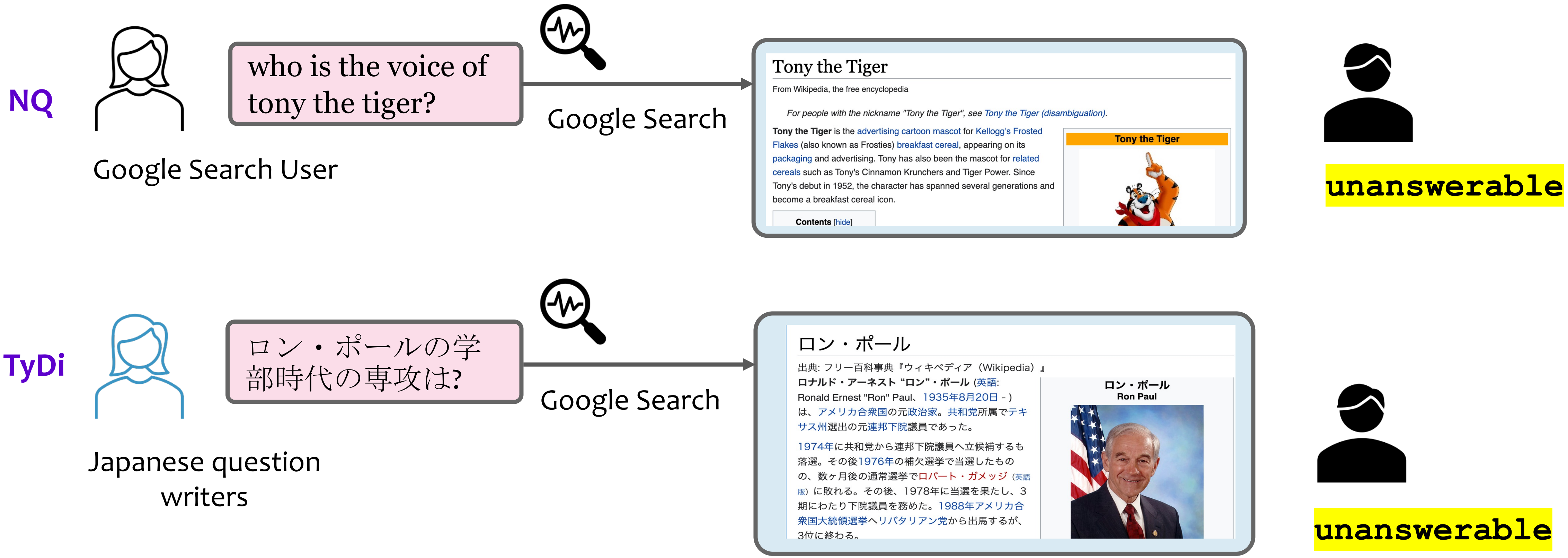


Annotating Unanswerability: Retriever failure

- Quality of the retriever (Search Engine): retrieved document is not most relevant.
- Missing information: No article with enough information in the target language.

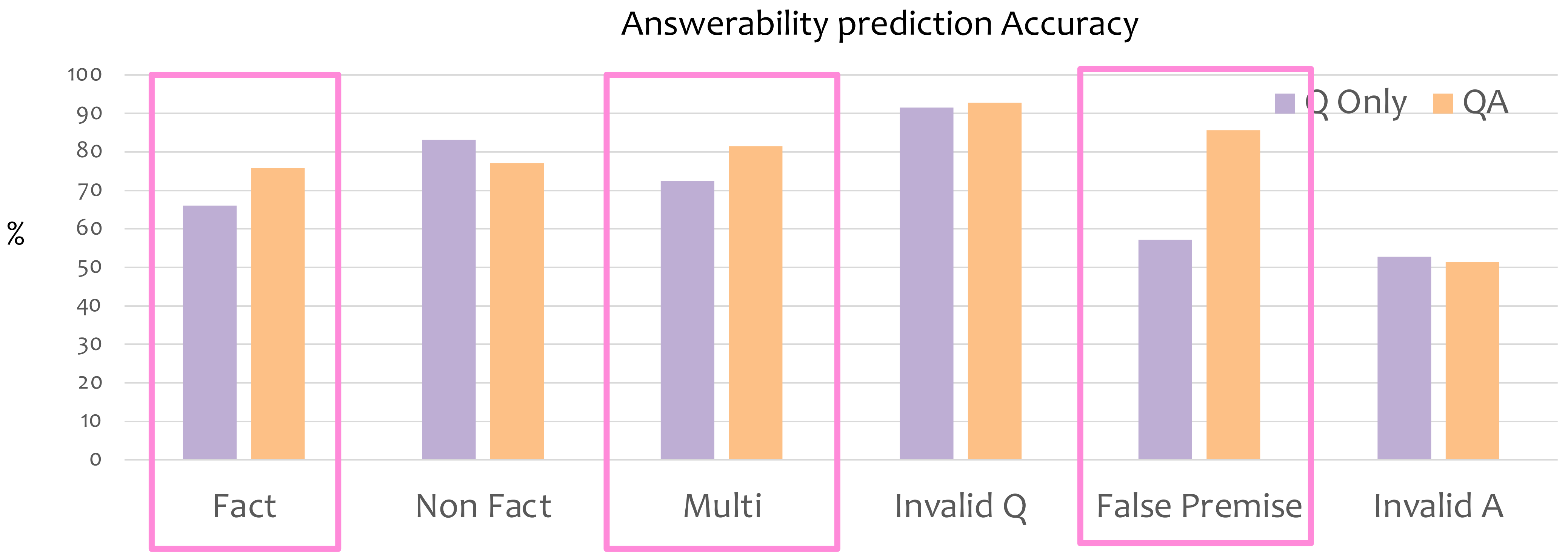
Annotating Unanswerability: Retriever failure

- Quality of the retriever (Search Engine): retrieved document is not most relevant.
- Missing information: No article with enough information in the target language.



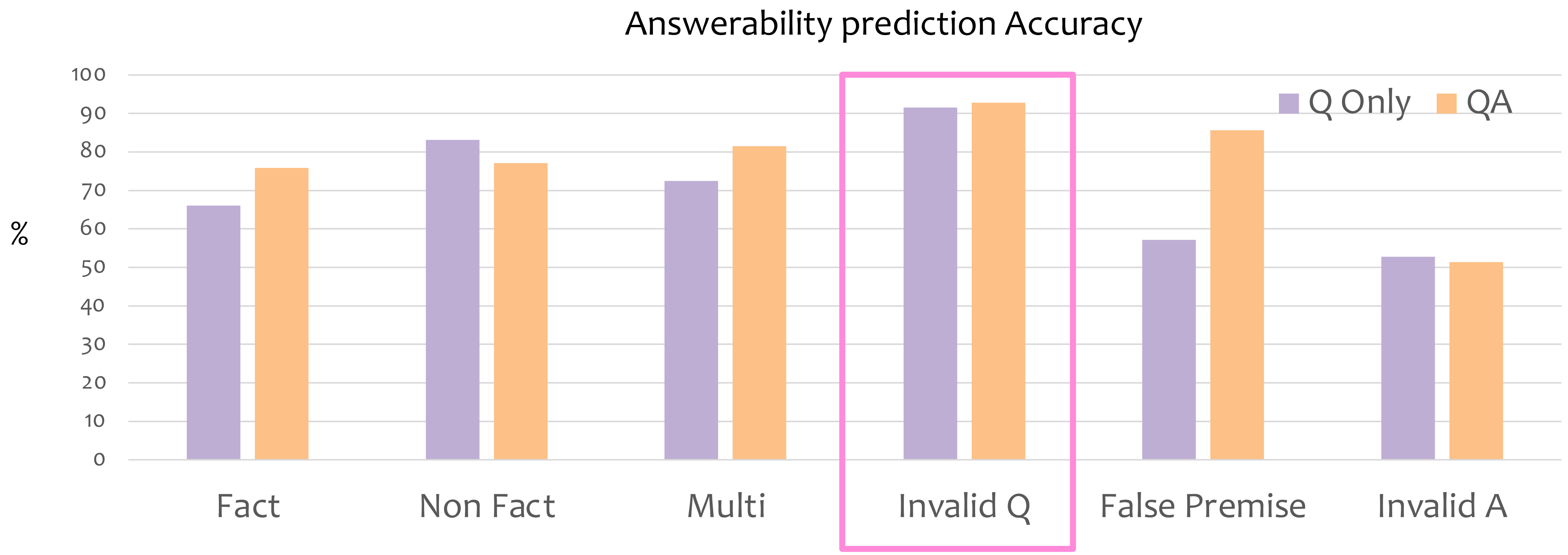
Annotating Unanswerability: per-category performance

- Context information helps models in **Factoid question**, **Multi-evidence question** and **False Premise.**



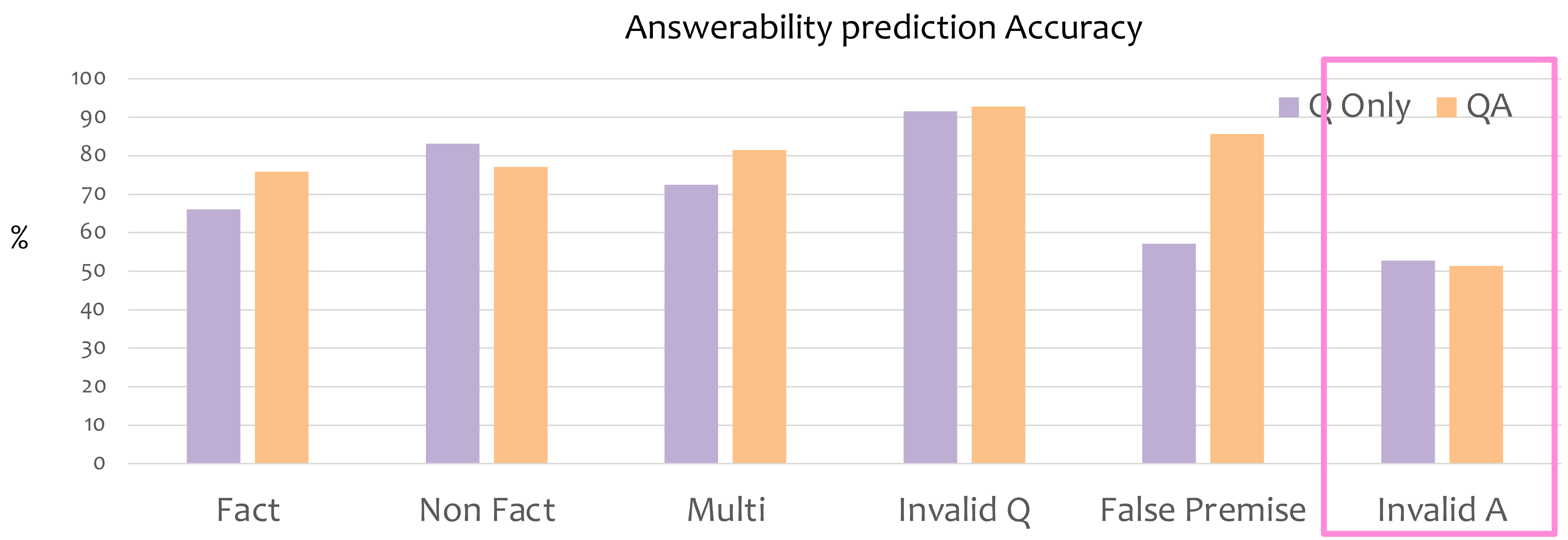
Annotating Unanswerability: per-category performance

- **Invalid question** can be easily detected even from question only.



Annotating Unanswerability: per-category performance

- **Invalid Answers** is hard to be predicted even given context input
- We found that our models often find answers overlooked by annotators.



Outlines

- **Background**
- **Gold Type / Gold Paragraph** experiments

  What are the remaining head-rooms in information-seeking QA?
- **Answerability Prediction** experiments

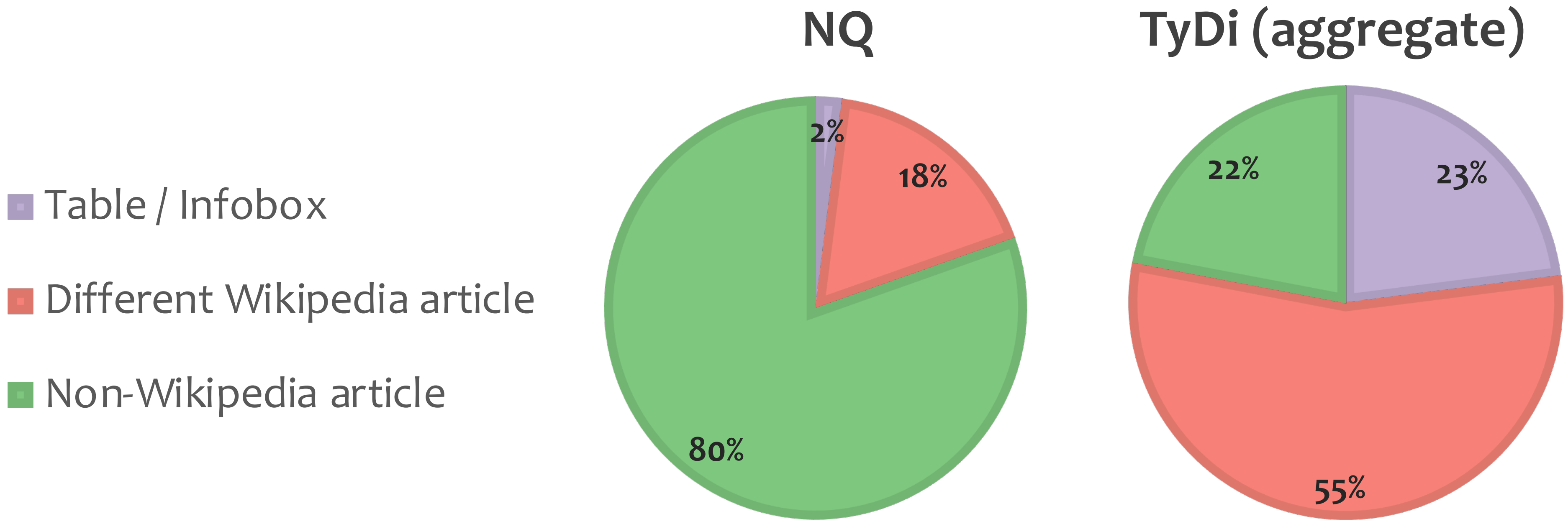
  How well do our models perform on answerability prediction?
- **Unanswerability Annotation**

  What makes the questions unanswerable?
- **Discussions**

  How can we improve answer coverage?

Discussions: Alternative knowledge sources

- We find that many retrieval miss questions can be answered if we use (i) data in structured format (e.g., Table, WikiData), (ii) non-Wikipedia articles.



- Reasoning across heterogeneous input remains challenging.

Discussions: New task formulations

Some questions can’t be answered by extracting from a single paragraph.

- **Generative QA**: Extracting answers may not sufficient.

what's written in the book at the end of it's a wonderful life

スペースシャトルと宇宙船の違いは何？
(What is the difference between a space shuttle and a spaceship?)

Discussions: New task formulations

Some questions can’t be answered by extracting from a single paragraph.

- **Generative QA**: Extracting answers may not sufficient.

> what's written in the book at the end of it's a wonderful life

> スペースシャトルと宇宙船の違いは何？
> (What is the difference between a space shuttle and a spaceship?)

- **Multi-evidence questions**: Single document may not entail questions.

> who was the king of england at the time the house of the seven gables was built

> 조선의 문신 출신 왕은 몇 명인가?
> (How many kings from Joseon dynasty used to be civil officers before becoming a king?)

Summary

We present comprehensive analysis on information-seeking QA dataset.

- Our controlled experiments show **Answerability Prediction** and **Paragraph Retrieval** remain challenging.
  - **Answerability Prediction:** models need to learn when to abstain from answering
  - **Paragraph Retrieval**: models have hard time picking the right paragraph.
- Our **unanswerability annotations** revel the unique cause of unanswerability and how we can improve future dataset development and task formulation.

Links

Paper ▶

Code & data (github) ▶