MeetingQA: Extractive Question-Answering on Meeting Transcripts

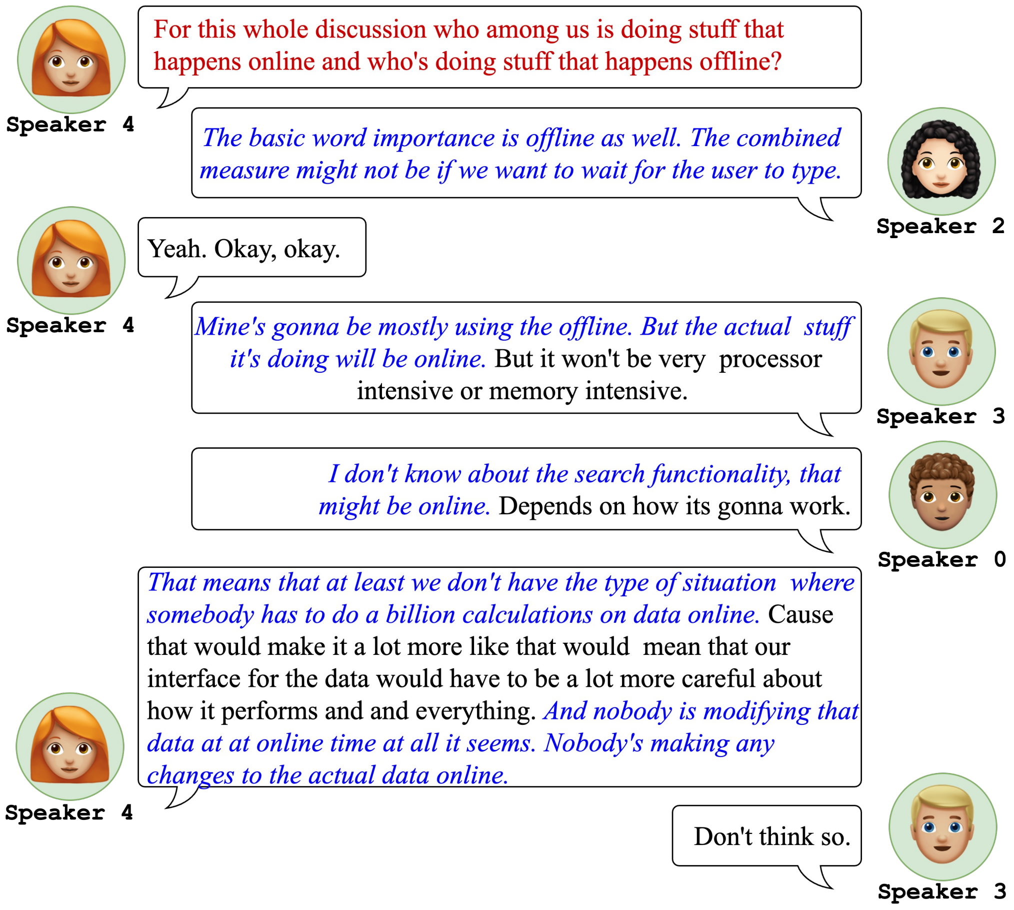
Archiki Prasad[1], Trung Bui[2], Seunghyun Yoon[2],
Hanieh Deilamsalehy[2], Franck Dernoncourt[2], Mohit Bansal[1]

[1]UNC Chapel Hill, [2]Adobe Research

Motivation

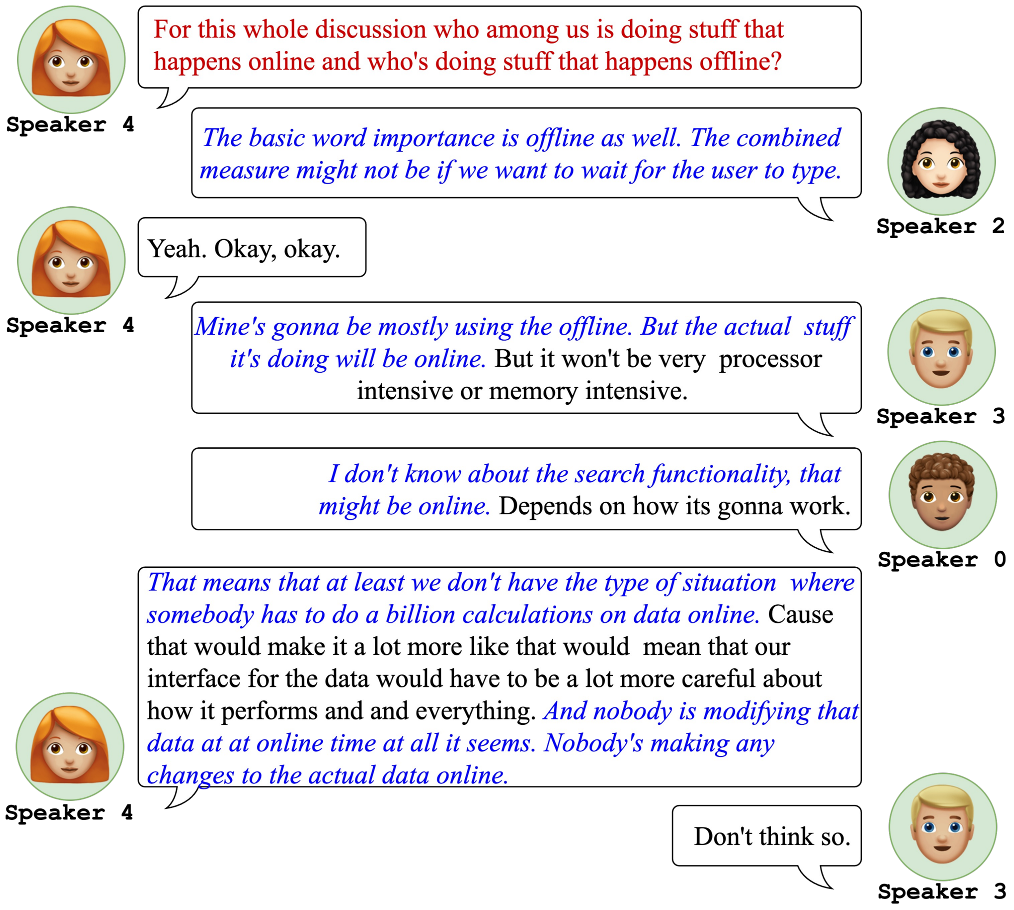
- Millions of meetings take place everyday worldwide
- Vast amounts of meeting transcripts
- What makes meeting transcripts unique?
  - Long documents
  - Domain-specific and information-rich
- Prior works focus on summarization and extracting action items
  - Under-utilize significant QA component in meeting discussions

MeetingQA: Introduction



- Extractive QA dataset based on *questions asked by participants in a meeting* and corresponding answer sentences

MeetingQA: Introduction



- Extractive QA dataset based on *questions asked by participants in a meeting* and corresponding answer sentences
- Why choose questions asked by participants?
  - Questions are longer, open-ended, and discussion-seeking
  - Rhetorical questions, multi-speaker, and multi-span answers

MeetingQA: Data Collection

Public transcripts from AMI corpus

~100 hours of manually transcribed multi-party meetings

Question Selection

Based on punctuation and length of question

Answer Annotation

Recruit annotators to label sentences in answer span

High inter-annotator agreement: Krippendorff's $\alpha = 0.73$

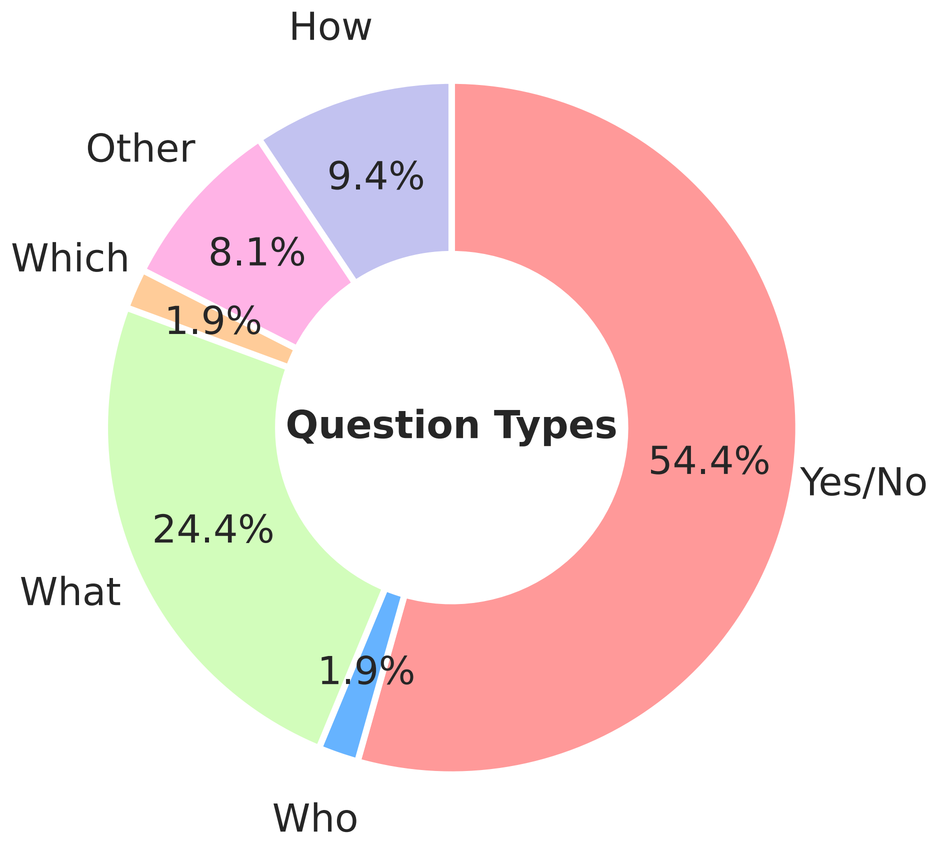
Carletta et al. "The AMI Meeting Corpus: a pre-announcement" in International Workshop on Machine Learning for Multimodal Interaction 2005

MeetingQA: Dataset Analysis

	Train	Dev	Test
Number of Meetings	64	48	54
Number of Questions	3007	2252	2476
w/ No Answer	956	621	764
w/ Multi-Span Answers	787	548	663
w/ Multi-Speaker Answers	1016	737	840
Avg. Questions per Meeting	46.98	46.92	45.85

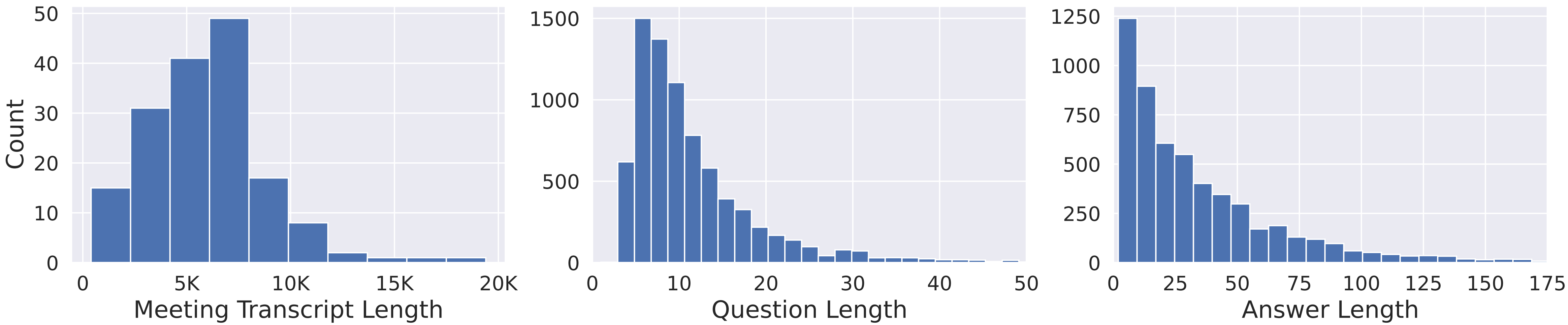
- Total of 7,735 questions from 166 different meetings split across train, dev, and test sets
- Statistics
  - Unanswerable Questions: 30%
  - Multi-span (non-consecutive sentences) answers: 40%
  - Multi-speaker answers: 48%

MeetingQA: Dataset Analysis



- Yes/no questions are information-seeking and detailed responses
- 50% questions are opinion-seeking
- 20% questions are framed rhetorically
- 70% of multi-speaker answers contain some disagreement

MeetingQA: Dataset Analysis



- Avg. Transcript: 5.9K words, Question: 12 words, and Answer: 35 words
- High human performance: F1 = 84.6

Methods

Context-retrieval for short-context models

Retrieve relevant segment of meeting transcript as context

Multi-span models

Using token classification models
- *I* tag: in answer span
- *O* tag: outside answer span

Single-span models

Single 'super' span: first to last relevant sentence in span

Silver data augmentation

Automatically annotated answer spans for questions from interviews in MediaSum dataset

Zhu et al. "MediaSum: A large-scale media interview dataset for dialogue summarization" in Proceedings of NAACL 2021

Experimental Results: Finetuned

	Model	Overall F1	No Ans. F1	Answerable F1		
				All	M-Span	M-Speaker
SS	RoBERTa-base	**56.5**	41.0	**63.1**	**60.8**	**64.1**
SS	Longformer-base	55.6	**46.1**	59.9	55.3	59.4
MS	RoBERTa-base	54.0	41.1	59.8	58.2	60.9
MS	Longformer-base	53.8	39.4	60.3	58.8	62.0
	Human Performance	**84.6**	**80.7**	**86.3**	**88.1**	**87.7**

Finetuned Performance

- ≥ 25 F1 points gap with human performance

Experimental Results: Finetuned

	Model	Overall F1	No Ans. F1	Answerable F1		
				All	M-Span	M-Speaker
SS	RoBERTa-base	**56.5**	41.0	**63.1**	**60.8**	**64.1**
SS	Longformer-base	55.6	**46.1**	59.9	55.3	59.4
MS	RoBERTa-base	54.0	41.1	59.8	58.2	60.9
MS	Longformer-base	53.8	39.4	60.3	58.8	62.0
	Human Performance	**84.6**	**80.7**	**86.3**	**88.1**	**87.7**

Finetuned Performance

- ≥ 25 F1 points gap with human performance
- Short-context models slightly outperform long-context

Experimental Results: Finetuned

	Model	Overall F1	No Ans. F1	Answerable F1		
				All	M-Span	M-Speaker
SS	RoBERTa-base	**56.5**	41.0	**63.1**	**60.8**	**64.1**
SS	Longformer-base	55.6	**46.1**	59.9	55.3	59.4
MS	RoBERTa-base	54.0	41.1	59.8	58.2	60.9
MS	Longformer-base	53.8	39.4	60.3	58.8	62.0
	Human Performance	**84.6**	**80.7**	**86.3**	**88.1**	**87.7**

Finetuned Performance

- ≥ 25 F1 points gap with human performance
- Short-context models slightly outperform long-context
- Multi-span models have slightly less or comparable performance than single-span models

Experimental Results: Zero-shot

Model	**Inter. Data**	**Overall F1**
RoBERTa-base	SQuADv2	27.9
	+ silver	**34.6**
Longformer-base	SQuADv2	15.1
	+ silver	32.5
FLAN-T5 XL	–	33.8
FLAN-T5 XL (self ans)	–	34.0
Human performance	–	**84.6**

Zero-shot Performance

- ~50 F1 points gap with respect to human performance

Experimental Results: Zero-shot

Model	Inter. Data	Overall F1
RoBERTa-base	SQuADv2	27.9
	+ silver	**34.6**
Longformer-base	SQuADv2	15.1
	+ silver	32.5
FLAN-T5 XL	–	33.8
FLAN-T5 XL (self ans)	–	34.0
Human performance	–	**84.6**

Zero-shot Performance

- ~50 F1 points gap with respect to human performance
- Silver data augmentation is effective

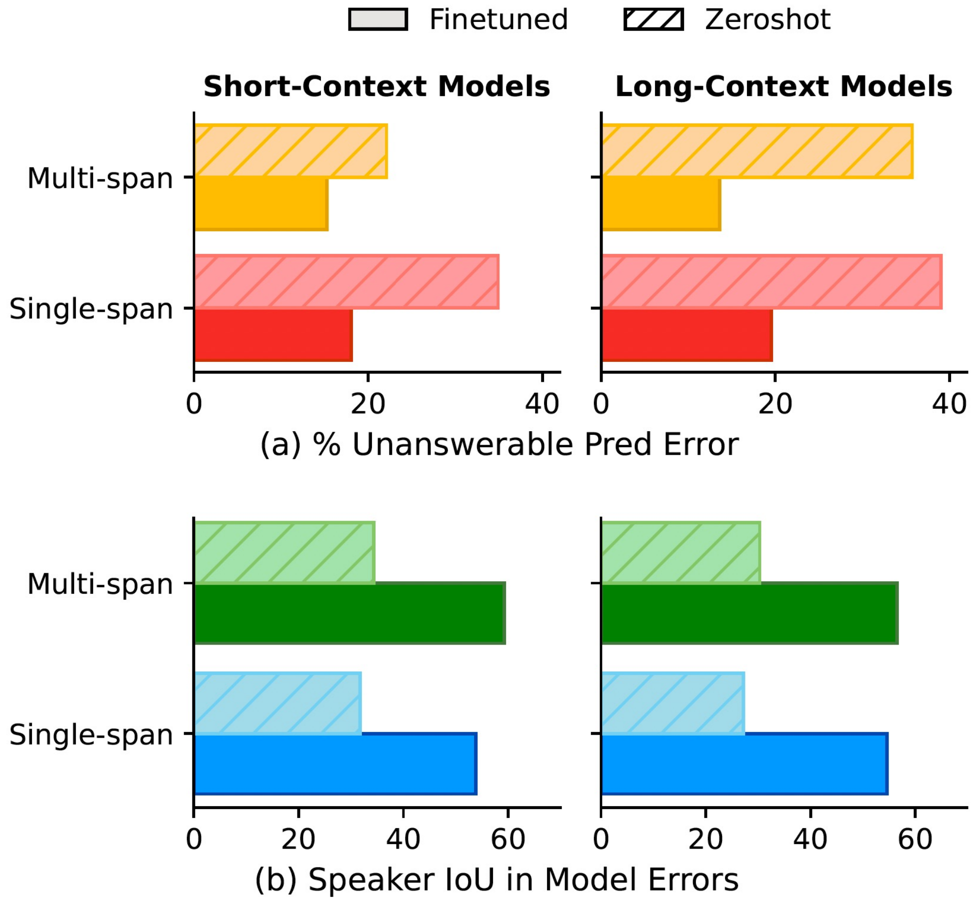
Experimental Results: Zero-shot

Model	Inter. Data	Overall F1
RoBERTa-base	SQuADv2	27.9
	+ silver	**34.6**
Longformer-base	SQuADv2	15.1
	+ silver	32.5
FLAN-T5 XL	–	33.8
FLAN-T5 XL (self ans)	–	34.0
Human performance	–	**84.6**

Zero-shot Performance

- ~50 F1 points gap with respect to human performance
- Silver data augmentation is effective
- Larger instruction tuned models yield comparable performance

Experimental Results: Error Analysis



- Models struggle at identifying rhetorical questions, especially in zero-shot setting
- Single-span predictions contain more irrelevant sentences
- Models struggle to identify which speakers answer a question, especially in zero-shot setting

Takeaways

- MeetingQA is an interesting QA dataset based on open-ended and discussion-heavy questions asked during meetings
- MeetingQA is challenging for existing QA models which lag behind human performance significantly
  - 25 F1 point gap in finetuned setting
  - 50 F1 point gap in zero-shot setting

Thank you for listening!

Project Page: https://archiki.github.io/meetingqa.html

Contact: archiki@cs.unc.edu