AmbigQA:
Answering Ambiguous Open-domain Questions

Sewon Min[1,2], Julian Michael[1], Hannaneh Hajishirzi[1,3], Luke Zettlemoyer[1,2]

[1]University of Washington, [2]Facebook AI Research, [3]Allen Institute of AI




facebook NLP Summit 2020, Question Answering Workshop

Motivation

When did Harry Potter and the Sorcerer's stone movie come out?

What season does Meredith and Derek get married in Grey's Anatomy?

Questions from NQ-open (Kwiakowski et al 2020)

Motivation

When did Harry Potter and the Sorcerer's stone movie come out?

Harry Potter and the Philosopher's Stone (film)

... had its world premiere at the Odeon Leicester Square in London on 4 November 2001 (...) released to cinemas in the United Kingdom and United States on 16 November 2001.

What season does Meredith and Derek get married in Grey's Anatomy?

Now or Never (Grey's Anatomy)

... Season 5 ... Meredith and Derek have decided not to wait any longer to get married and just go to City Hall that evening ... writes their vows down on a post-it note that they both sign ...

Grey's Anatomy (Season 7)

... She and Derek decide to adopt Zola, an orphaned baby, and make their marriage legal.

Motivation

When did Harry Potter and the Sorcerer's stone movie come out?

Harry Potter and the Philosopher's Stone (film)

... had its world premiere at the Odeon Leicester Square in London on 4 November 2001 (...) released to cinemas in the Kingdom and United ... 01.

What season ... get married ... 's Anatomy?

Over 50% of questions from NQ are ambiguous

Now or Never (Grey's Anatomy)

... Season 5 ... Meredith and Derek have decided not to wait any longer to get married and just go to City Hall that evening ... writes their vows down on a post-it note that they both sign ...

Grey's Anatomy (Season 7)

... She and Derek decide to adopt Zola, an orphaned baby, and make their marriage legal.

Motivation

When did Harry Potter and the Sorcerer's stone movie come out?

Harry Potter and the Philosopher's Stone (film)

... had its world premiere at the Odeon Leicester Square in London on 4 November 2001 (...) released to cinemas in [illegible] Kingdom and

Wha[illegible]

cided

ust go to

down

on a post-it note that they both sign ...

Grey's Anatomy (Season 7)

... She and Derek decide to adopt Zola, an orphaned baby, and make their marriage legal.

Open-domain questions are inherently ambiguous

When people ask questions in new / unfamiliar topics, questions cannot be guaranteed to have a single clear answer (even when people intend to do so)

AmbigQA task – option 1

When did Harry Potter and the Sorcerer's stone movie come out?

4 November 2001

16 November 2001

Less helpful for users due to lack of context

What season does Meredith and Derek get married in Grey's Anatomy?

Season 5

Season 7

AmbigQA task – option 2

When did Harry Potter and the Sorcerer's stone movie come out?

The movie's first premiere at the Odeon Leicester Square was on 4 November 2001 and then the movie was released to cinem… November 2001 …

Less well-defined, Non-trivial evaluation, Cannot separate answer prediction vs. disambiguation

What season does Meredith and Derek get married in Grey's Anatomy?

Meredith and Derek got informally married with a post-it note in Season 5 and then they later made it legal when they adopt a baby in Season 7…

AmbigQA task – our definition

When did Harry Potter and the Sorcerer's stone movie come out?

Q: When did harry potter and the sorcerer's stone movie come out at the Odeon Leicester Square?
A: 4 November 2001

Q: When did harry potter and the sorcerer's stone movie come out in cinemas?
A: 16 November 2001

What season does Meredith and Derek get married in Grey's Anatomy?

Q: What season does Meredith and Derek get informally married in Grey's Anatomy?
A: Season 5

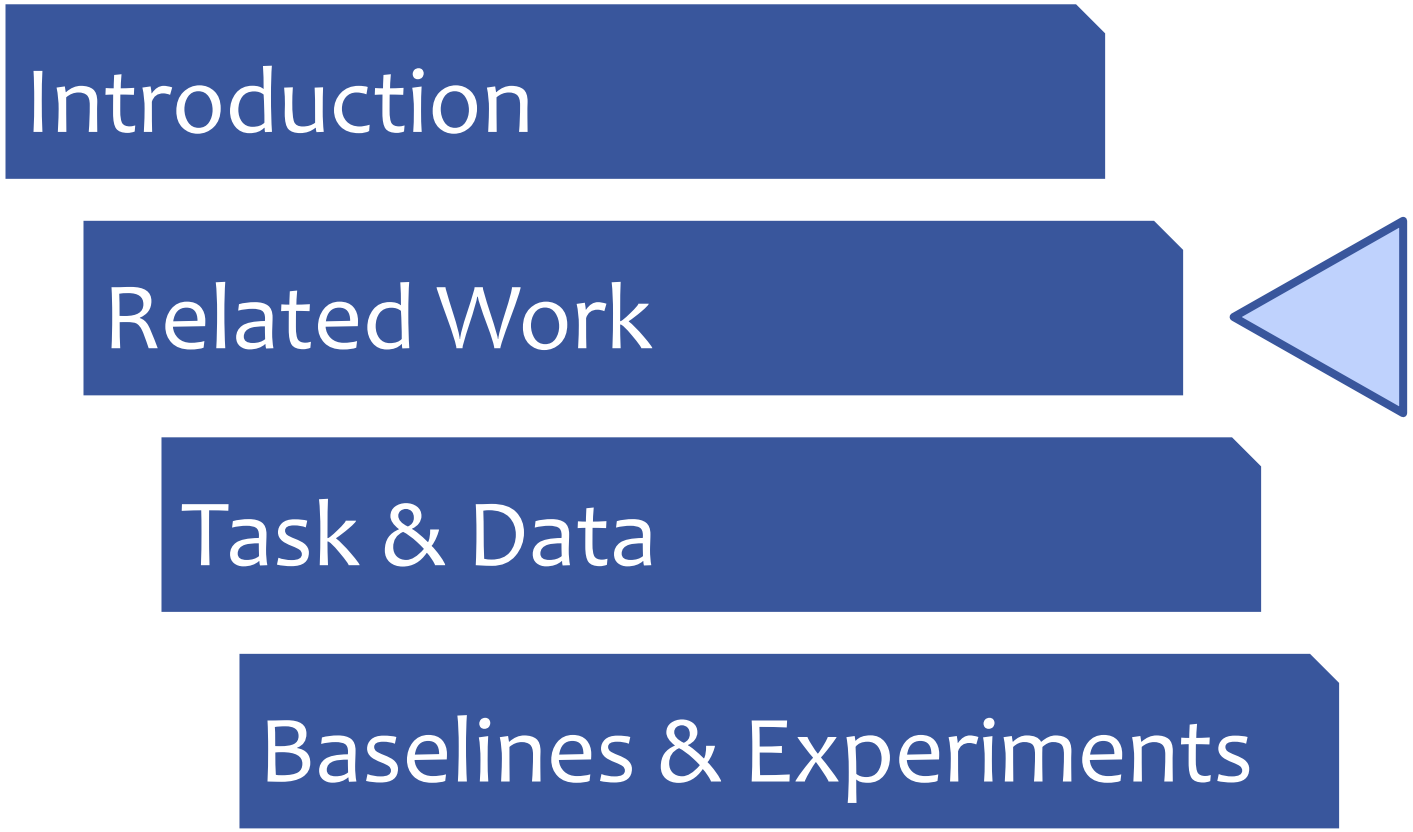
Q: What season does Meredith and Derek get legally married in Grey's Anatomy?
A: Season 7

Explicit answers to the original question
+ disambiguation in a more well-defined way

Contribution

- We introduce **AmbigQA**, a new task that answers to the open-domain questions by identifying all plausible answers along with their disambiguations.
- We construct a **dataset** with 14,042 annotations on NQ-open questions containing ambiguity that is frequent, diverse & subtle.
- We introduce the **first baseline models**, with experiments showing their effectiveness in learning from our data while highlighting avenues for future work.

Content


Introduction
Related Work
Task & Data
Baselines & Experiments

Related work (1/2)

Open-domain question answering

- Long-standing problem in NLP
- Questions vary from information-seeking (Berant et al 2013, Kwiatkowski et al 2019, Clark et al 2019) to more specialized trivia/quiz (Joshi et al 2017, Dunn et al 2017)
- Assume each question has a single clear answer
- Nonetheless, Kwiatkowski et al 2019 report that the answers are often debatable; an average pairwise agreement of the answers in Natural Questions annotations is 49.2%

Related work (1/2)

Open-domain question answering

- Long-standing problem in NLP
- Questions vary from information-seeking (Berant et al 2013, Kwiatkowski et al 2019, Clark et al 2019) to more specialized trivia/quiz (Joshi et al 2017, Dunn et al 2017)
- Assume each question has a single clear answer
- Nonetheless, Kwiatkowski et al 2019 report that the answers are often debatable; an average pairwise agreement of the answers in Natural Questions annotations is 49.2%

We **embrace ambiguity** as **inherent** to information-seeking questions

Related work (2/2)

Asking clarification questions

- Questions that are annotated by crowdworkers (Xu et al 2019)
- Simple, vague keywords, e.g. *"dinasour"* (Zhai et al 2003, Aliannejadi et al 2019)

Related work (2/2)

Asking clarification questions

- Questions that are annotated by crowdworkers (Xu et al 2019)
- Simple, vague keywords, e.g. *"dinasour"* (Zhai et al 2003, Aliannejadi et al 2019)

We study **unintentional** & **subtle** ambiguity in natural questions

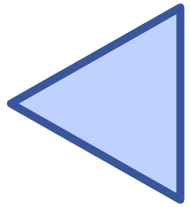
We provide **complete** & **immediate** solution
(necessary as users often do not know which information they want before having access to all answers & their context)

Content

Introduction

Related Work

Task & Data



Baselines & Experiments

Task Definition

- Given a prompt question q, output n question-answer pairs $\{(x_i, y_i)\}_{i=1}^{n}$ where each y_i is an equally plausible answer to q, and each x_i is a minimal modification of q whose answer is unambiguously y_i
- If q has a single answer, output y_1

Task Definition

- Given a prompt question q, output n question-answer pairs $\{(x_i, y_i)\}_{i=1}^{n}$ where each y_i is an equally plausible answer to q, and each x_i is a minimal modification of q whose answer is unambiguously y_i
- If q has a single answer, output y_1

Evaluation metrics

- Multiple Answer Prediction (F_{answer})
- Full task (F_{BLEU}, $F_{\text{EDIT-F1}}$): Consider a similarity score between reference and generated question (BLEU or EDIT-F1)

Task Definition

- Given a prompt question q, output n question-answer pairs $\{(x_i, y_i)\}_{i=1}^{n}$ where each y_i is an equally plausible answer to q, and each x_i is a minimal modification of q whose answer is unambiguously y_i
- If q has a single answer, output y_1

Evaluation metrics

- Multiple Answer Prediction (F_{answer})
- Full task (F_{BLEU}, $F_{\text{EDIT-F1}}$): Consider a similarity score between reference and generated question (**BLEU** or EDIT-F1)

Task Definition

- Given a prompt question q, output n question-answer pairs $\{(x_i, y_i)\}_{i=1}^{n}$ where each y_i is an equally plausible answer to q, and each x_i is a minimal modification of q whose answer is unambiguously y_i
- If q has a single answer, output y_1

Evaluation metrics

- Multiple Answer Prediction (F_{answer})
- Full task (F_{BLEU}, $F_{\text{EDIT-F1}}$): Consider a similarity score between reference and generated question (BLEU or **EDIT-F1**)

Prompt	Who made the play the crucible?
Reference	Who wrote the play the crucible?
Generated	Who made the play the crucible in 2012?

-made, +wrote

+in, +2012

EDIT-F1: 0

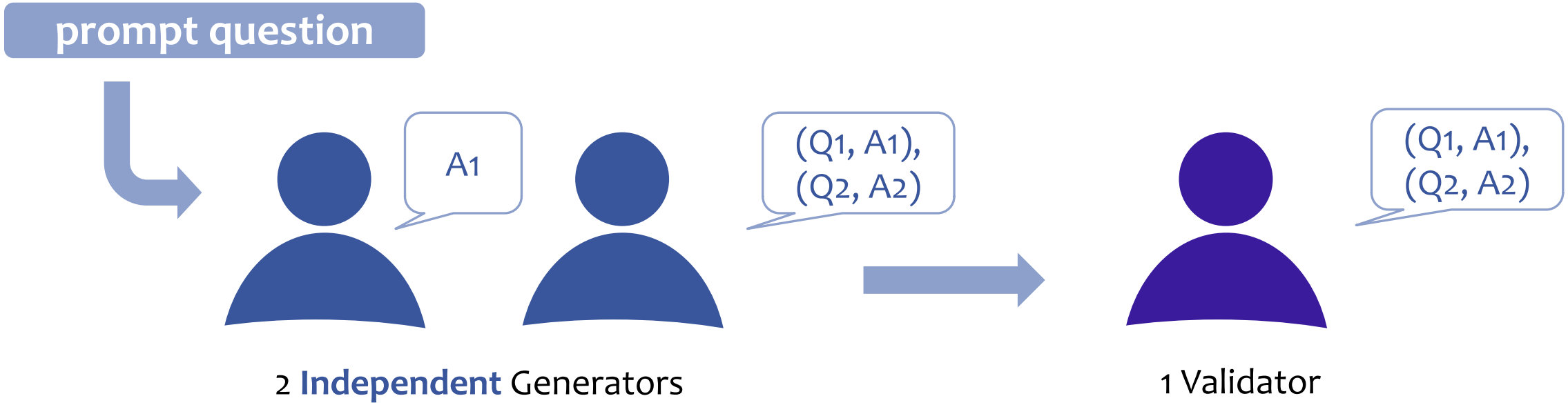
Data Collection

- Maximizing recall is difficult even for humans, as ambiguity is not easily found.
- We were able to collect high quality data with high levels of ambiguity using **careful worker selection** and **an annotation pipeline: generation + validation**

Data Collection

- Maximizing recall is difficult even for humans, as ambiguity is not easily found.
- We were able to collect high quality data with high levels of ambiguity using **careful worker selection** and **an annotation pipeline: generation + validation**

Information-seeking questions from NQ (Kwiatkowski et al 2019)



Data Analysis

- 14,042 questions
- Over 50% of questions are ambiguous
- Diverse types of ambiguity with fairly long-tailed edits

Type of Ambiguity	%
Event references	39
Properties	27
Entity references	23
Answer types	16
Time-dependency	13
Multiple sub-questions	3

(c) Word cloud of the edits made in questions; ■ and ■ indicate added and deleted unigrams, respectively.

Data Analysis

- 14,042 questions
- Over 50% of questions are ambiguous
- Diverse types of ambiguity with fairly long-tailed edits

Type of Ambiguity	%
Event reference	
Prop	
Entity references	23
Answer types	16
Time-dependency	13
Multiple sub-questions	3

More analyses in the paper

(c) Word cloud of the edits made in questions; ■ and ■ indicate added and deleted unigrams, respectively.

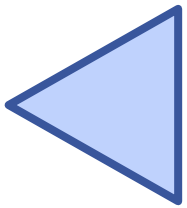
Content

Introduction

Related Work

Task & Data

Baselines & Experiments



Please see the paper for details!

Baselines

Step 1: Multi-answer prediction

DPR (Karpukhin et al 2020)

BART (Lewis et al 2020)

(1) Thresholding over likelihood of each span

(2) **SpanSeqGen:** Generate a sequence of answers, separated by [SEP]

Step 2: Question disambiguation

Prompt question
Targeted answer
Untargeted answers
Passages

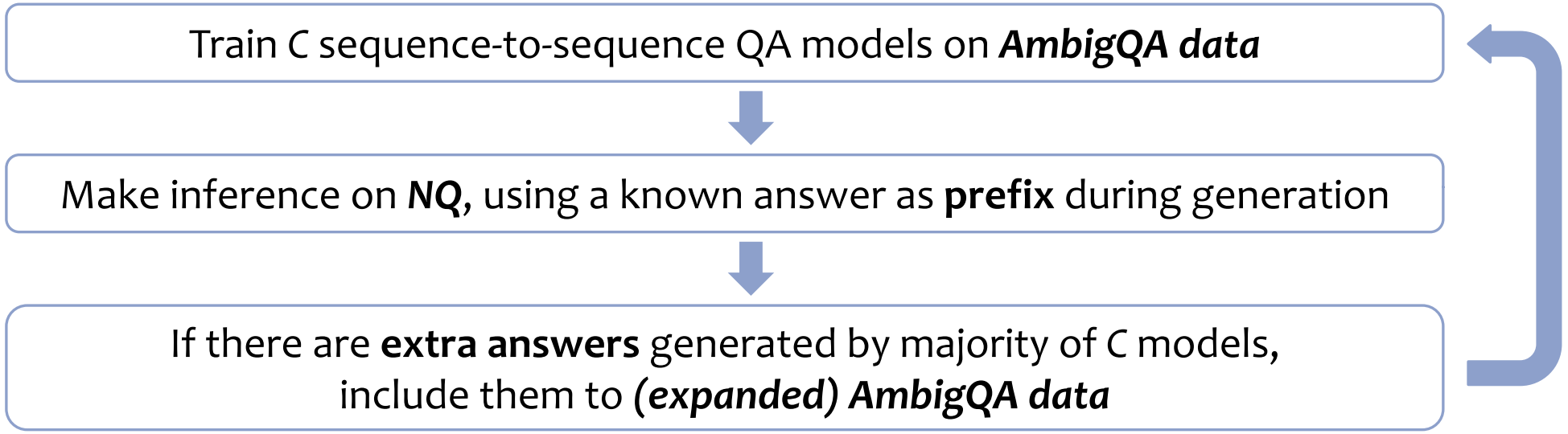
→ BART →

Edited question

Please see the paper for details!

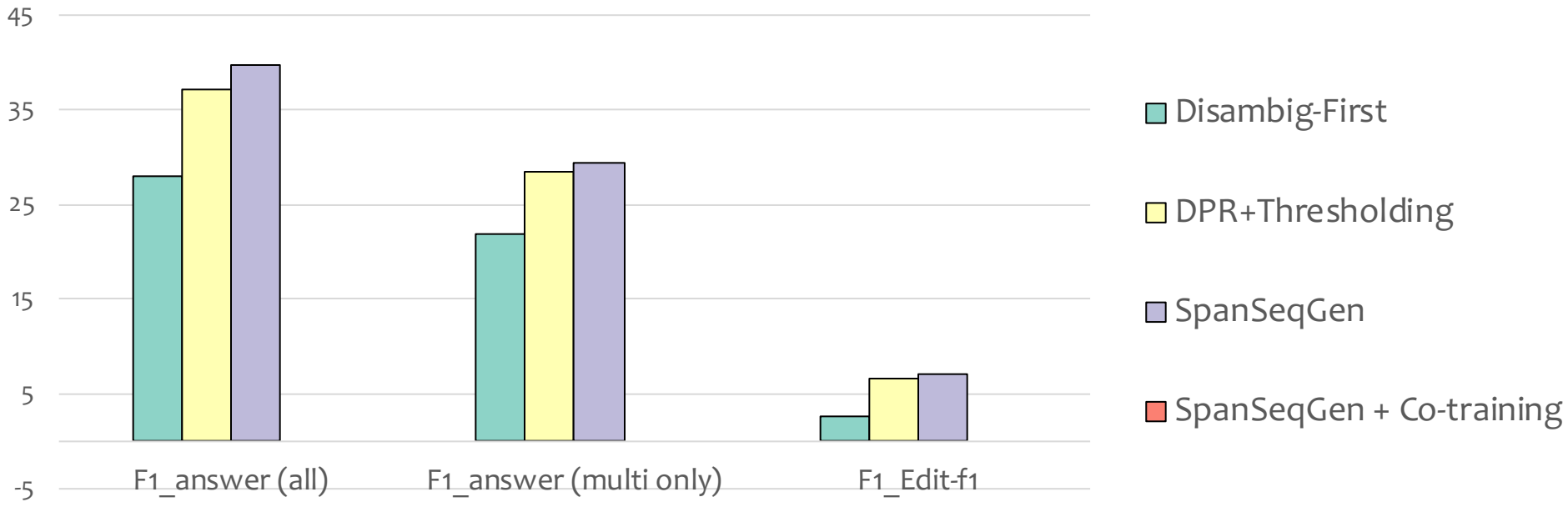
Modified Democratic Co-training (Zhou & Goldman, 2004)

- Prevalence of unlabeled ambiguity in NQ-open → Let's use ***a single known answer*** from NQ-open as ***weak supervision***



Please see the paper for details!

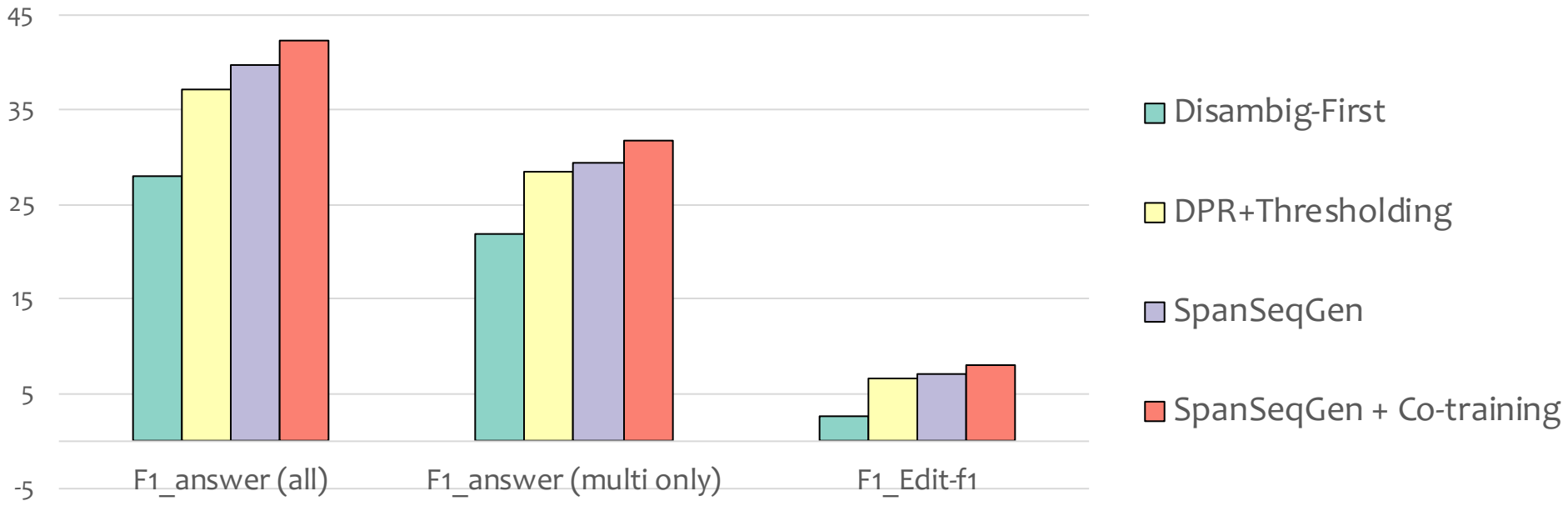
Results



SpanSeqGen outperforms other baselines.

(Esp. Disambig-First, which does disambiguation before reading passages)

Results

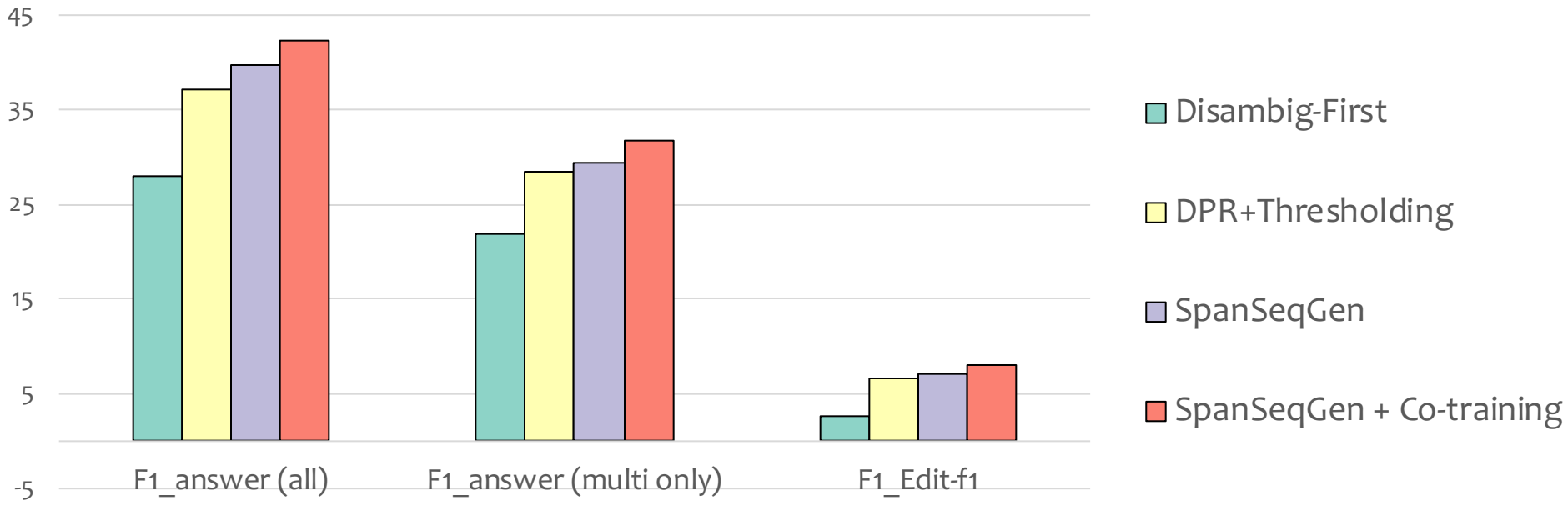


SpanSeqGen outperforms other baselines.

(Esp. Disambig-First, which does disambiguation before reading passages)

SpanSeqGen + Co-training further boosts the performance

Results



SpanSeqGen outperforms other baselines.

(Esp. Disambig-First, which does disambiguation before reading passages)

SpanSeqGen + Co-training further boosts the performance

Still huge room for improvements; **more analyses in the paper!**

Thank you!

Paper: https://arxiv.org/abs/2004.10645

Website (+Data): https://nlp.cs.washington.edu/ambigqa/

Eval script (+ baseline codes): https://github.com/shmsw25/AmbigQA

facebook research