Answering Ambiguous Questions through Generative Evidence Fusion and Round-Trip Prediction

Yifan Gao

The Chinese University of Hong Kong

Henghui Zhu, Patrick Ng, Cicero Nogueira dos Santos, Zhiguo Wang, Feng Nan, Dejiao Zhang, Ramesh Nallapati, Andrew O. Arnold, Bing Xiang

AWS AI

Motivation

When people ask information-seeking questions,
they do not have the knowledge of relevant topics.

⇩

It can be difficult to formulate clear and unambiguous questions.
>50% Google search queries are ambiguous! (Min et al., 2020)

An Example in Open-Domain Question Answering

Q: What's the most points scored in an NBA game?

Wikipedia
The Free Encyclopedia

Search within Wikipedia

Relevant Wikipedia Passage 1:
The highest-scoring regular season game is the triple-overtime game between ... the two teams combined to score **370** points, with the pistons defeating the nuggets **186**–184 ...

Relevant Wikipedia Passage 2:
Wilt Chamberlain scored an nba-record **100** points ...

You may want to ask:

- What's the most points scored in an NBA game by combined team? / 370
- What's the most points scored in an NBA game by a single team? / 186
- What's the most points scored in an NBA game by an individual? / 100

AmbigQA Task

- > 50% questions in NQ-Open QA dataset (Kwiatkowski et al. 2019) are ambiguous
- AmbigQA: A new open-domain QA dataset which involves **disambiguating** and **answering** a potentially ambiguous question (Min et al., 2020)

Given:

- A prompt question q
- The whole collection of Wikipedia passages (25 Million!)

- Answer Prediction Subtask: find one or multiple answers $a_1, \ldots, a_n$
- Question Disambiguation Subtask: if multiple answers are predicted ($n > 1$), rewrite the prompt question q into ***disambiguated*** q_i for a_i

Challenges

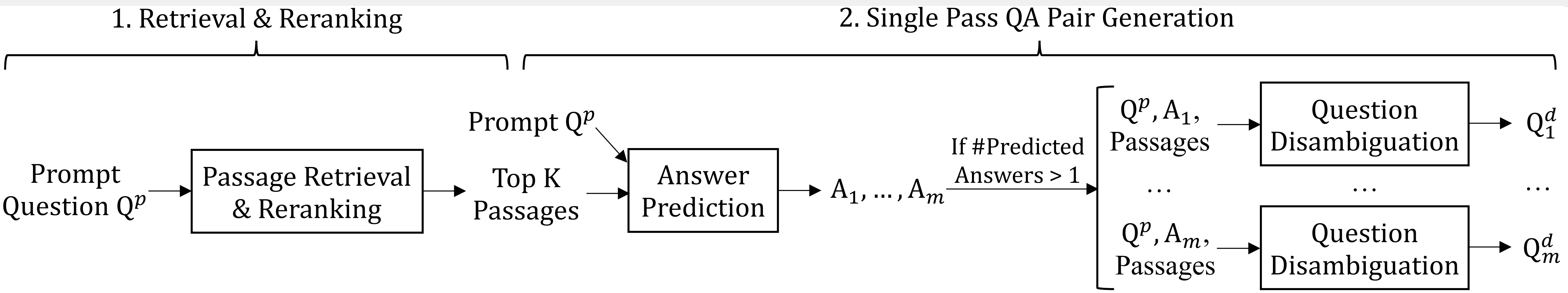
- Existing methods cannot find all interpretations for disambiguation
  - SpanSeqGen (previous best model): 1.17 QA pairs per question
  - Ground Truth: 2.19 QA pairs per question
  - SpanSeqGen can encode ~8 passages at most
- Mismatch between Question Generation pretraining and Question Disambiguation finetuning
  - Question Generation (NQ-open): Answer + Passage -> Question
  - Question Disambiguation (AmbigNQ): Answer + Question + Passage -> Disambiguated Question

REFUEL: Round-trip Evidence FUsion via gEneration with retrieval

- Find more interpretations per ambiguous question (1.72 QA pairs, ↑47%)
  - Round-trip generation with conditional-probability-based filtering
  - Board coverage of knowledge passages via Fusion-in-Decoder (100 passages)

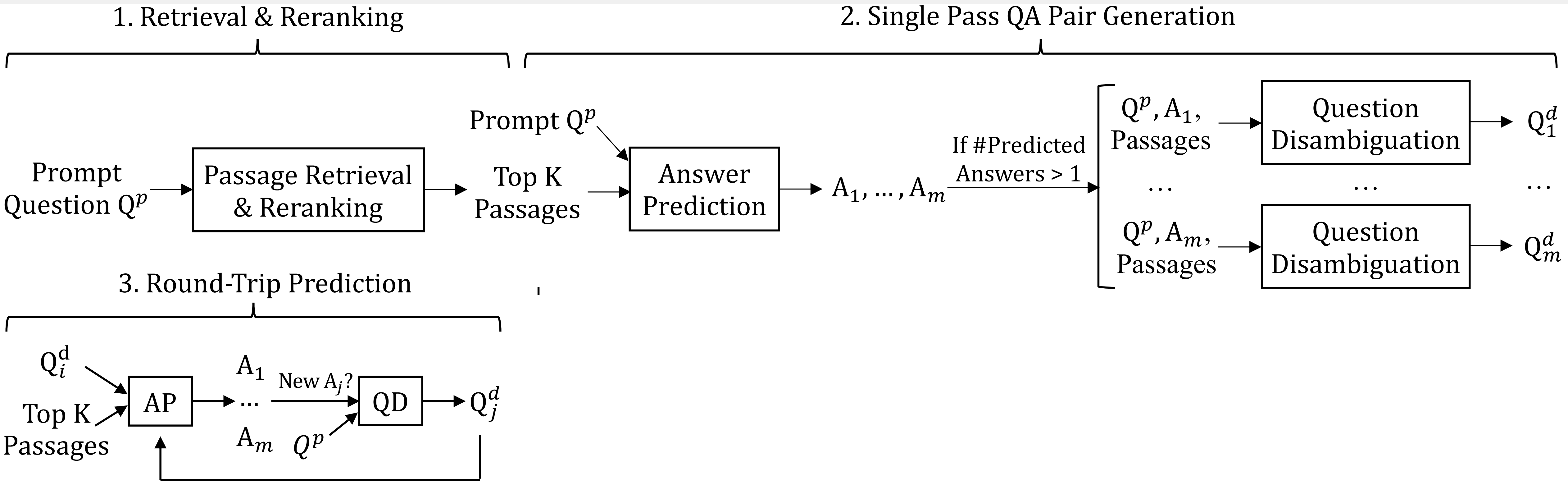
- Improvement on question disambiguation
  - Token-Deletion Pretraining
  - Insertion-based weighted loss

REFUEL: A General Framework



1. Dense Passage Retrieval + BERT Reranking ⟶ 1000 Wikipedia passages
2. A single pass QA pair generation model (architecture agnostic) makes the first prediction pass ⟶ a set of QA pairs

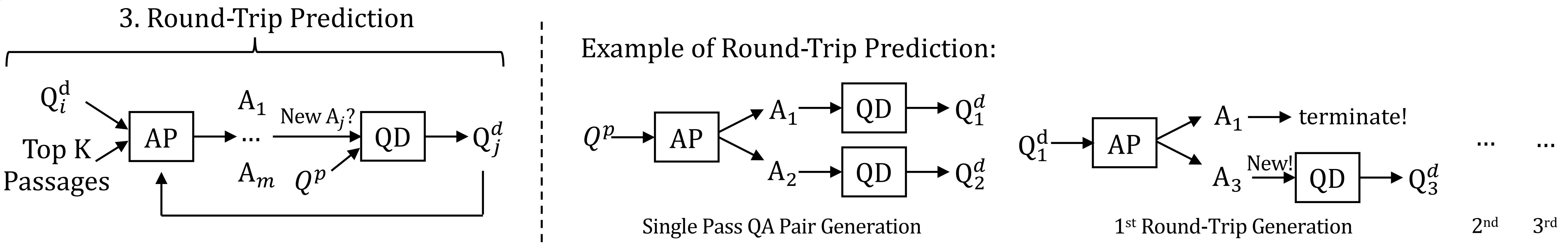
REFUEL: A General Framework



3. Round-Trip Prediction

- Round-Trip Generation: Find more interpretations missed in the first prediction pass
- Language Model Verification: Refine using a conditional-probability-based filtering approach

Round-Trip Prediction

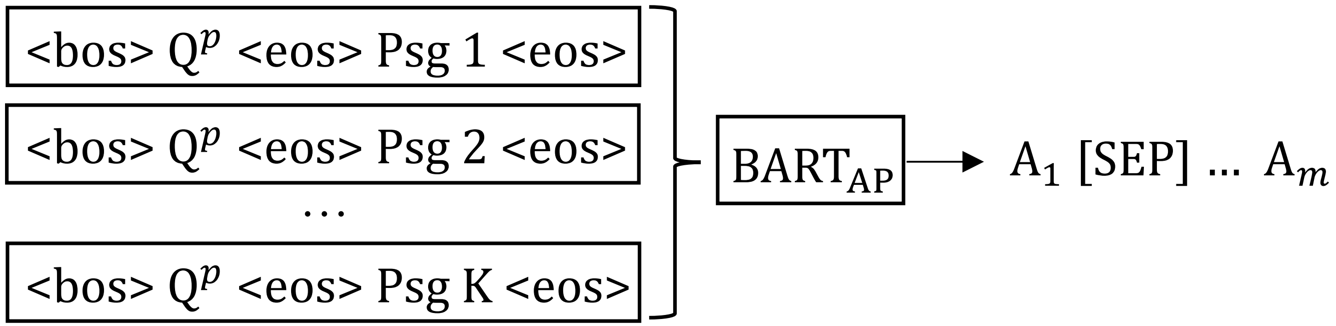


- Round-Trip Generation
  - Take the generated disambiguated questions as input
  - Harvest more answer candidates
- Language Model (LM) Filtering
  - Train an open-domain QA model on unambiguous QA pairs
  - Filtering according to the LM score: likelihood of the predicted answer given the question and passages

$$\text{LM score} = \sum_{i=1}^{N_a} \log p(a^i | q, \text{Psg})$$

Our Single Pass QA Model: Answer Prediction

- Baseline:
  - $\mathrm{Q} + \mathrm{Psg}_1 + \mathrm{Psg}_2 + \ldots + \mathrm{Psg}_\mathrm{T} -> \mathrm{A}_1$ [SEP] $\ldots \mathrm{A}_\mathrm{m}$ ($\mathrm{T} \leq 8$ passages)
- Our Approach:
  - Process each passages individually in $\mathrm{BART}_\mathrm{AP}$ encoder
  - $\mathrm{BART}_\mathrm{AP}$ decoder performs attention over them to fusion evidence
  - Scale up to 100 passages to find more interpretations



- Pretrain: NQ-open (80k samples); Finetune: AmbigNQ (10k samples)

Our Single Pass QA Model: Question Disambiguation

- How to effectively leverage NQ-open (80k) for pretraining?

- Naïve Approach: train a question generation model
  - Answer + Passage -> Question
- Mismatch between pretraining and finetuning!
  - Pretrain on NQ-open (80k): Answer + Passage -> Question
  - Finetune on AmbigNQ (10k): Answer + Prompt Q + Passage -> Disambiguated Q

Our Single Pass QA Model: Question Disambiguation

Token-Deletion Pretraining: construct "*ambiguous*" questions in NQ-open

1. Randomly delete an informative span within the question:
   - Informative span: the span containing at least one of the following Part-of-Speech tags: 'ADJ', 'NOUN', 'NUM', 'PROPN', 'SYM', 'VERB'.
2. Learn to recover key information from passages to rewrite the partial question into the complete question

Pretraining: Partial Question → Complete Question

When does the come out? + June 20, 2011 + Passages → *When does the Fifty Shades of Grey come out?*

Aligned!

Finetuning: Prompt Question → Disambiguated Question

When does the Fifty Shades of Grey come out? + June 20, 2011 + Passages → *When did movie the Fifty Shades of Grey come out in Los Angeles?*

Our Single Pass QA Model: Question Disambiguation

- Challenge: Disambiguated Q is very similar to the prompt Q
  - Directly copy the prompt questions as predictions can achieve 47.4 BLEU!

- **Insertion-based Weighted Loss**: put more emphasis on the newly added tokens of the disambiguated questions

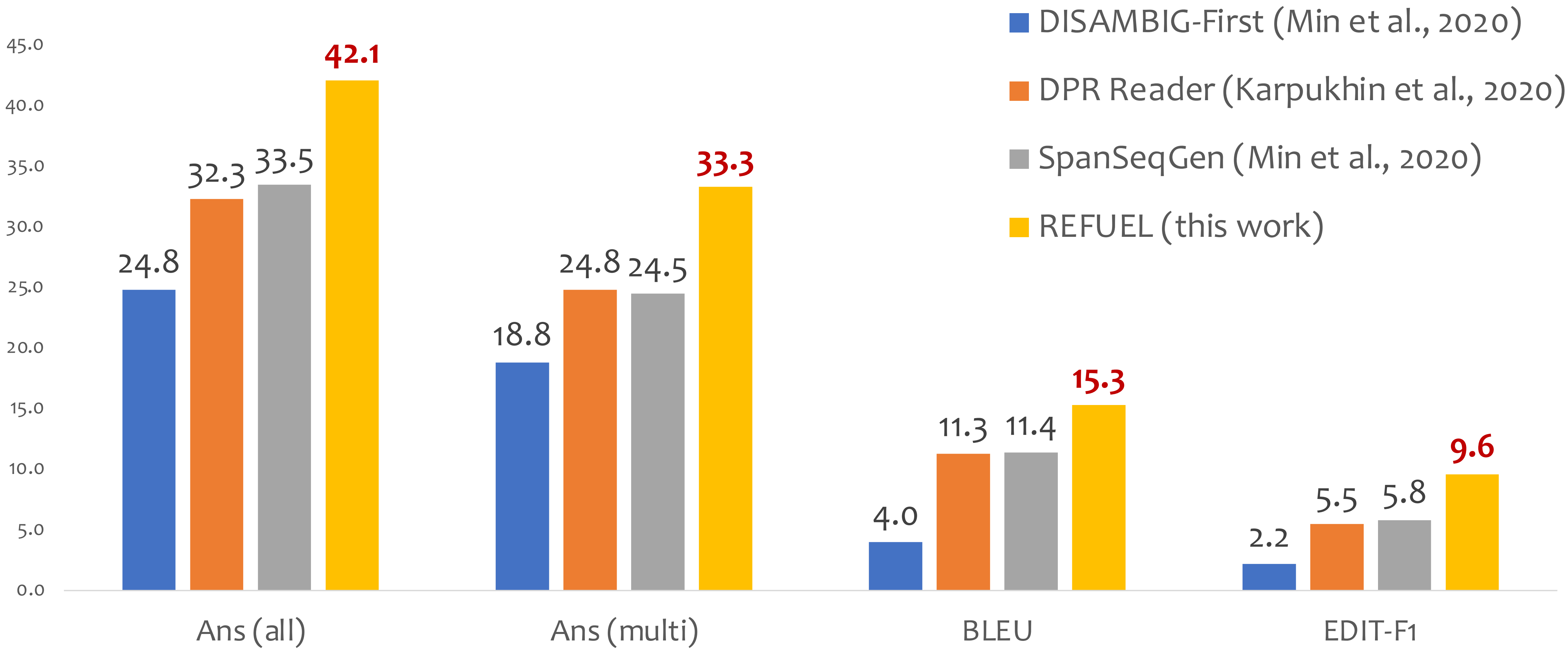
$$\mathcal{L} = \mathcal{L}_{nll} - \lambda \sum_{q_j \in \{q^{in}\}} \log(q_j | A, Q^p, \mathrm{Psg})$$

Prompt Question		Disambiguated Question
When does the Fifty Shades of Grey come out?	→	*When did movie the Fifty Shades of Grey come out in Los Angeles?*

Dataset & Evaluation Metrics

- Dataset: AmbigNQ (Min et al., 2020)
  - >50% questions are ambiguous, 2.1 distinct answers per question
  - Train / Dev / Test (not public) dataset sizes are 10036 / 2002 / 2004
  - Pretraining: Natural Questions (NQ-open) : 80k questions
- Evaluation Metrics
  - Answer Prediction: F-score between gold & predicted answers
    - $F1_{ans}$ (all): All questions
    - $F1_{ans}$ (multi): subset of questions which have multiple answers
  - Question Disambiguation:
    - BLEU: BLEU score between prompt and disambiguated Q ☹
    - EDIT-F1: focus on the added / deleted unigrams from the prompt to the disambiguated Q ☺

Leaderboard Performance (Hidden Test Set)



Effect of Round-Trip Prediction

Round-Trip Prediction (RTP) is a model-agnostic general approach: it can help our REFUEL as well as baselines!

Models	#QAs	Ans (multi)	EDIT-F1
REFUEL (w/o RTP)	1.55	37.0	11.2
REFUEL	1.72	37.4	11.8
DPR Reader	1.62	29.9	6.8
DPR Reader + RTP	1.81	31.6	7.3
SpanSeqGen	1.14	29.3	7.1
SpanSeqGen + RTP	1.28	29.9	7.4

Human Evaluation Results

3 workers evaluate the correctness of disambiguated QA pairs

Let $(q_1\ a_1), (q_2\ a_2)\ ...\ (q_n\ a_n)$ be the n generated QA pairs from the same prompt question:

- #C-QAs: (q_i, a_i) is considered correct if a_i is a correct answer of q_i
- #CD-QAs: (q_i, a_i) is considered correct iff:
  1) a_i is a correct answer of q_i
  2) any a_j (j ≠ i) is a wrong answer of q_i

Models	Dataset	#QAs	#C-QAs	#CD-QAs	κ
SPANSEQGEN	AMBIGQA	2.12	1.40	0.46	0.27
REFUEL w/o RTP	AMBIGQA	2.80	1.84	0.98	0.35
REFUEL	AMBIGQA	3.44	2.40	1.24	0.34
REFUEL w/o RTP	NQ-OPEN	2.32	1.30	0.64	0.20
REFUEL	NQ-OPEN	3.20	1.72	0.88	0.21
REFUEL w/o RTP	TriviaQA	2.08	1.02	0.46	0.34
REFUEL	TriviaQA	3.24	1.84	0.82	0.35

113%↑

Round-Trip Prediction (RTP) can find more correct interpretations!

Conclusion

- A general approach for answering ambiguous questions
  - Round-Trip Generation
  - Language Model Verification
- A better single pass QA model:
  - Scalable Retrieval-Augmented Generation for Evidence Fusion
  - Question Disambiguation
    - Token-deletion Pretraining Task
    - Insertion-based Weighted Loss
- Code & Models: https://github.com/amzn/refuel-open-domain-qa