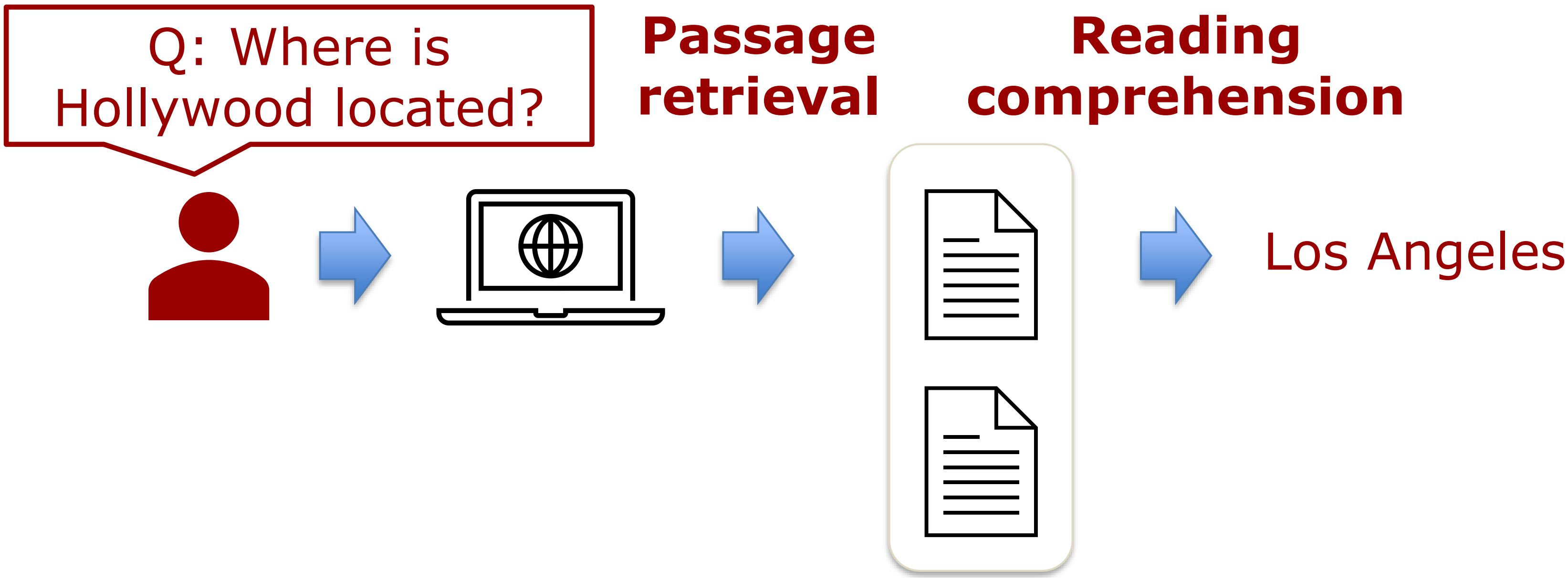
Answer Consolidation: Formulation and Benchmarking

Wenxuan Zhou[1], Qiang Ning[2], Heba Elfardy[2], Kevin Small[2], Muhao Chen[1]
University of Southern California[1], Amazon[2]

Open-domain Question Answering

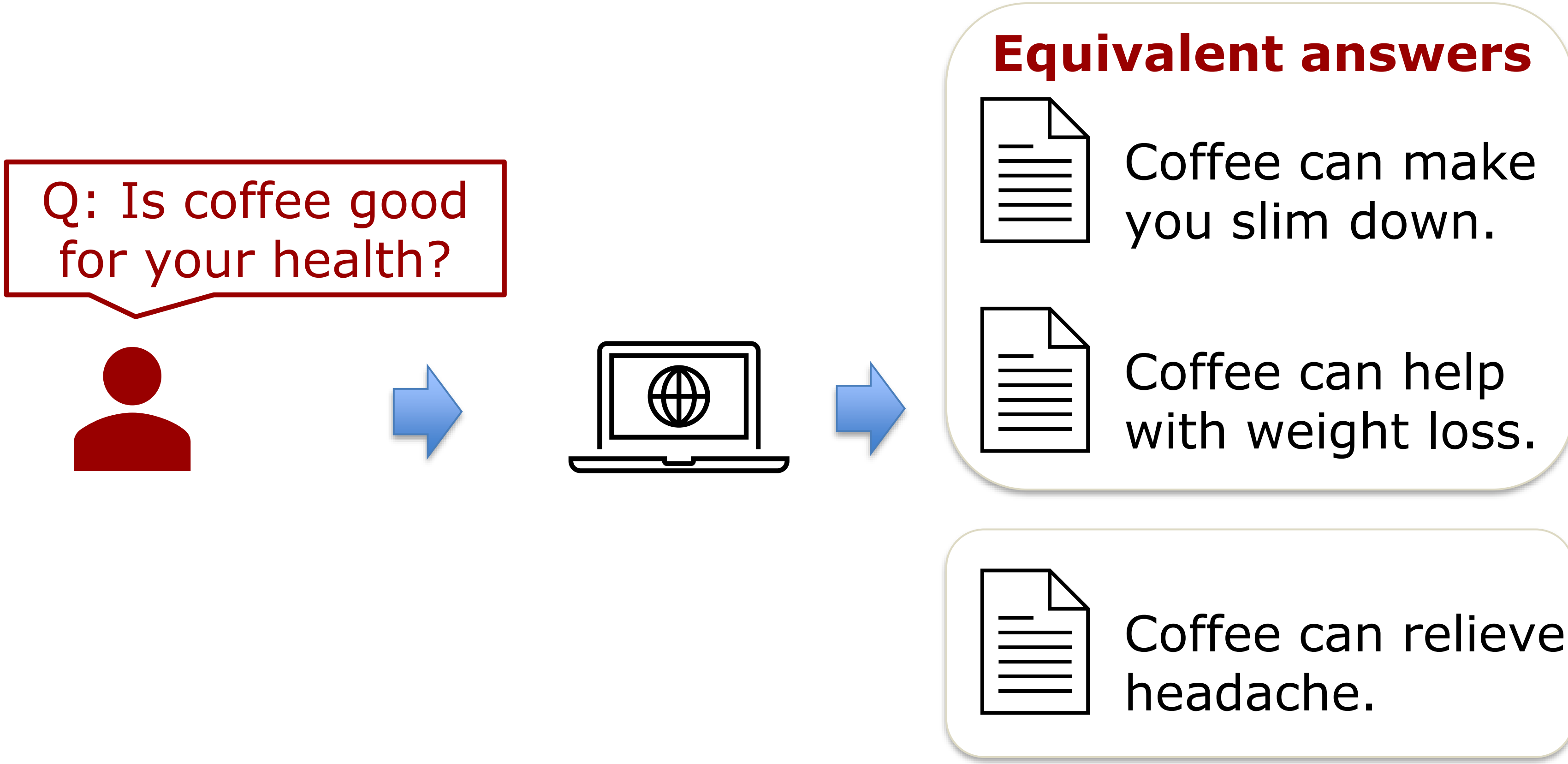
Answering natural language questions using large collections of documents.



Datasets: Open-domain SQuAD, Natural Questions, …

Single-answer assumption

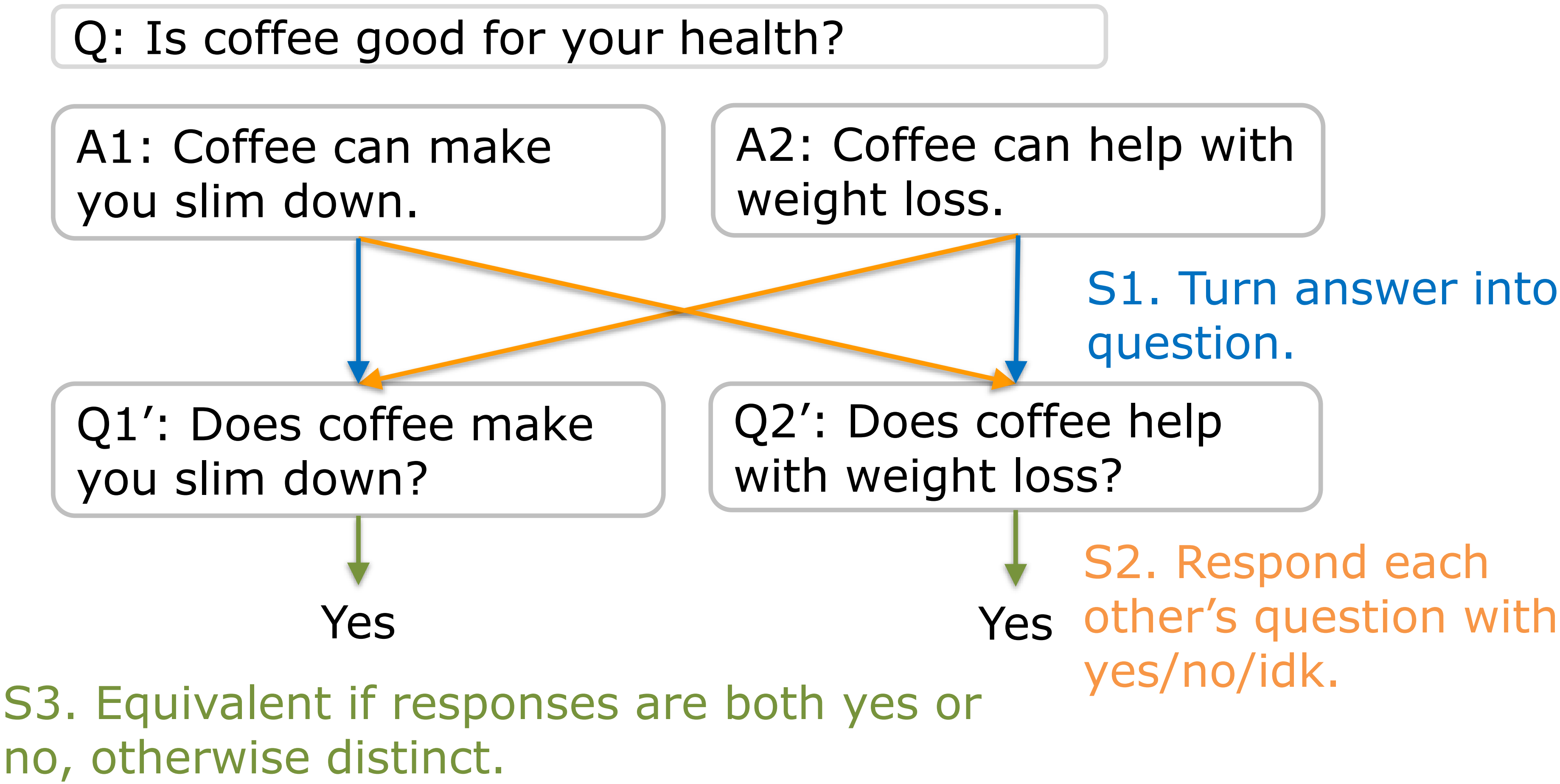
Multi-answer Scenario



Goal: group equivalent answers, identify distinct answers.

Problem Formulation

Equivalent/distinct answers: answers containing same/different perspectives, opinions, angles, etc.

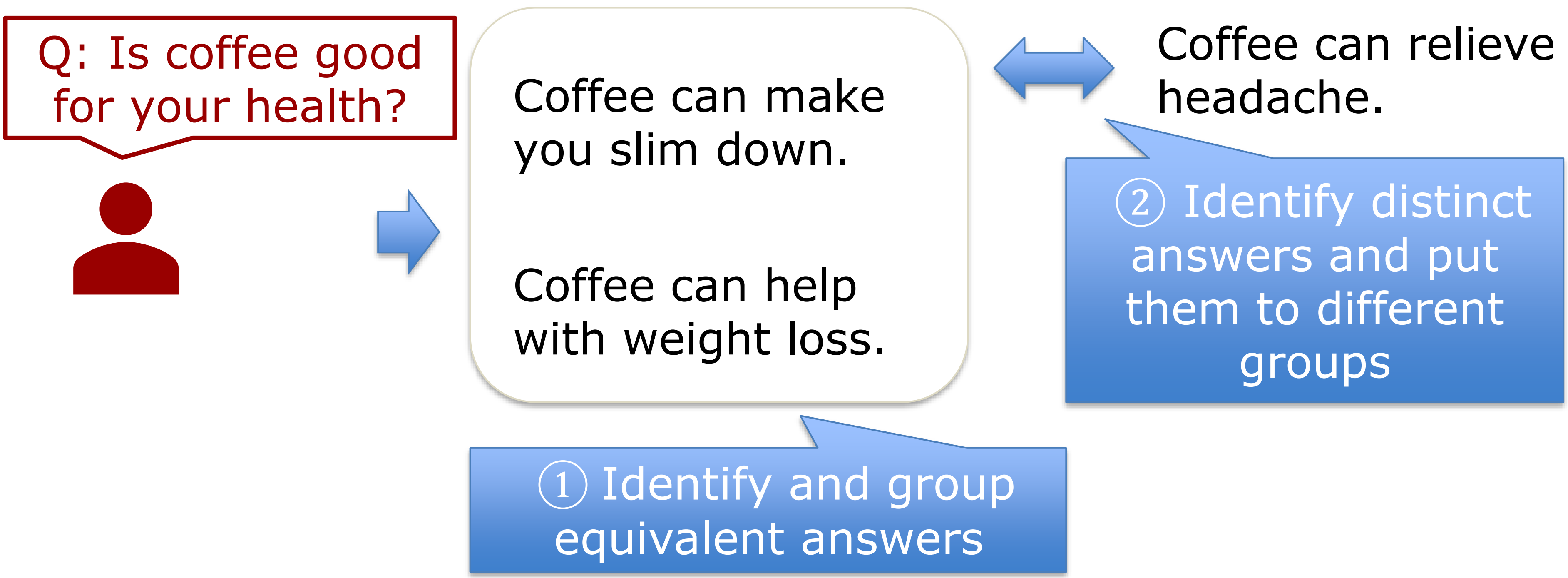


Problem Formulation (Cont.)

Task: Given a question and some answers, put them into groups such that:

(1) each answer belongs to one group.

(2) answers from the same/different groups are equivalent/distinct.



QUASI Dataset: Construction

S1. Quora questions (QQP) → **S2. Sentence Retrieval** → **Answers (in form of sentences)**

↓

S3. MTurk

Q: Is coffee good for your health?

1. Coffee can help you burn fat.
2. Drinking warm water can help you relax.

…

11. Coffee can cause insomnia and restlessness.

Add group | Remove empty groups

Sentence groups:

Not an answer:

Hard to put into groups:

← Groups of equivalent answers

QUASI Dataset: Statistics

4,699 questions, 24,006 answers, 19,676 groups. Train: Dev: Test = 80%: 10%: 10%.

Types of equivalent answers:

1. Exact match (56%)
2. Lexical variation (11%)
3. Semantic variation (30%)

Q: How does the respiratory system work?
S1: The respiratory system works by <u>getting the good air in and</u> <u>the bad air out</u>.
S2: The Respiratory System a simple system designed to <u>get oxygen into the body, and to get rid of carbon dioxide and water</u>.

4. Ambiguous (3%; Wrong annotation)

Experiments: Evaluation Settings

1. Sentence pair classification

- Given a question and two answers, decide whether they are equivalent.

2. Sentence grouping

- Put answers into groups.

Both under zero-shot and supervised settings.

Experiments: Models

Bi-encoders:

- Inputs:

$$\text{<s>}X_q\ X_s\text{</s>}$$

- Prediction: cosine similarity

Cross-encoders

- Inputs:

$$\text{<s>}X_q\ X_{s_1}\text{</s></s>}X_q\ X_{s_2}\text{</s>}$$

- Prediction: linear classifier

Answer-aware cross-encoders

- Inputs: extract the answer spans and add to inputs

$$\text{<s>}X_q\ X_{s_1}\ X_{a_1}\text{</s></s>}X_q\ X_{s_2}\ X_{a_2}\text{</s>}$$

- Prediction: linear classifier

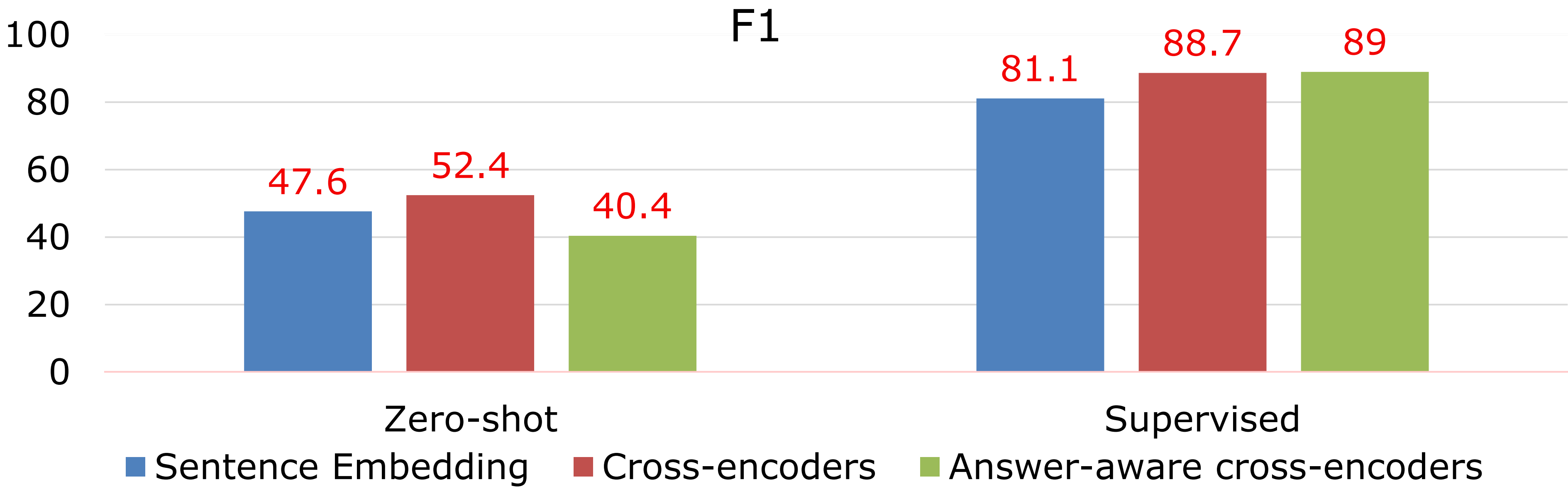
X_q: question

X_s: sentence

X_a: extracted answer

Experiments: Main results

Encoders: SimCSE-RoBERTa for bi-encoder, RoBERTa-MNLI for cross-encoder.
Setting: sentence pair classification.



The best supervised model achieves ~90% F1

Experiments: Error Analysis

Randomly sample 50 equivalent answers that are mistakenly classified as distinct:

1. Exact match (2%)

   - Estimated recall: 99.5% ⇒ easy to identify

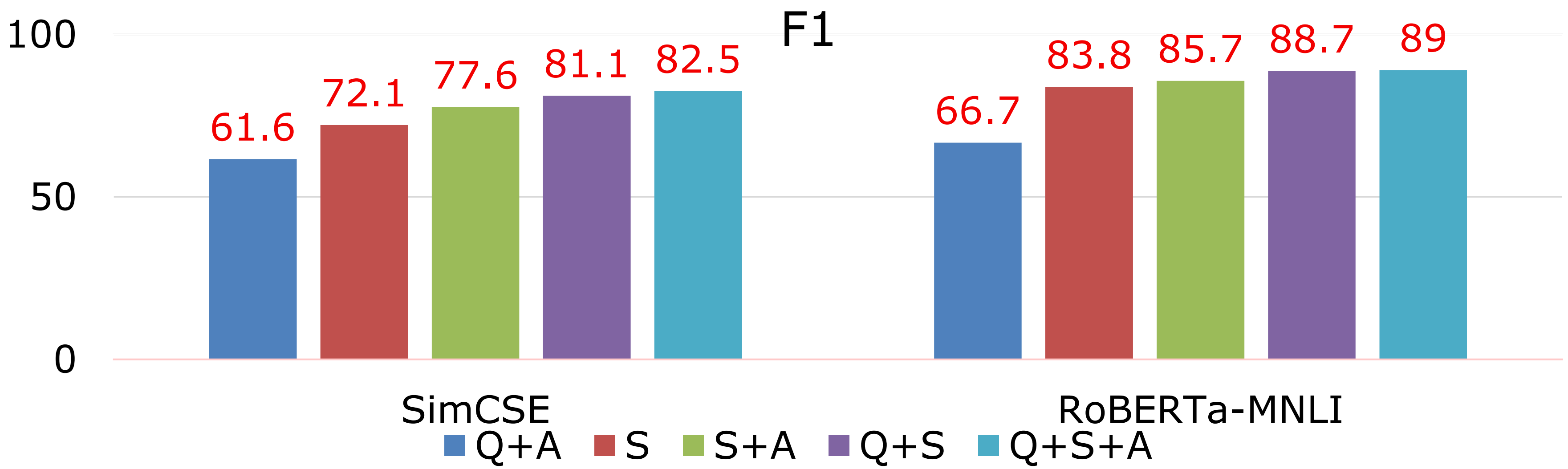
2. Ambiguous (16%)

3. Semantic variations (82%)

   - Estimated recall: 66.7% ⇒ large room for improvements

Experiments: Ablation

Q: question
S: sentences
A: answer spans extracted by UnifiedQA



Observations:
1. Removing S ⇒ largest drop.
2. Removing Q and A ⇒ 2nd largest drop.

Conclusion

1. We formulate and propose **answer consolidation**.

2. We contribute the **Question-Answer consolidation dataset** (QUASI) and benchmark with various types of methods.

3. Experiments suggest room for further studies on more **robust and generalizable solutions** for answer consolidation, which would benefit real-world QA systems.

Code & Data