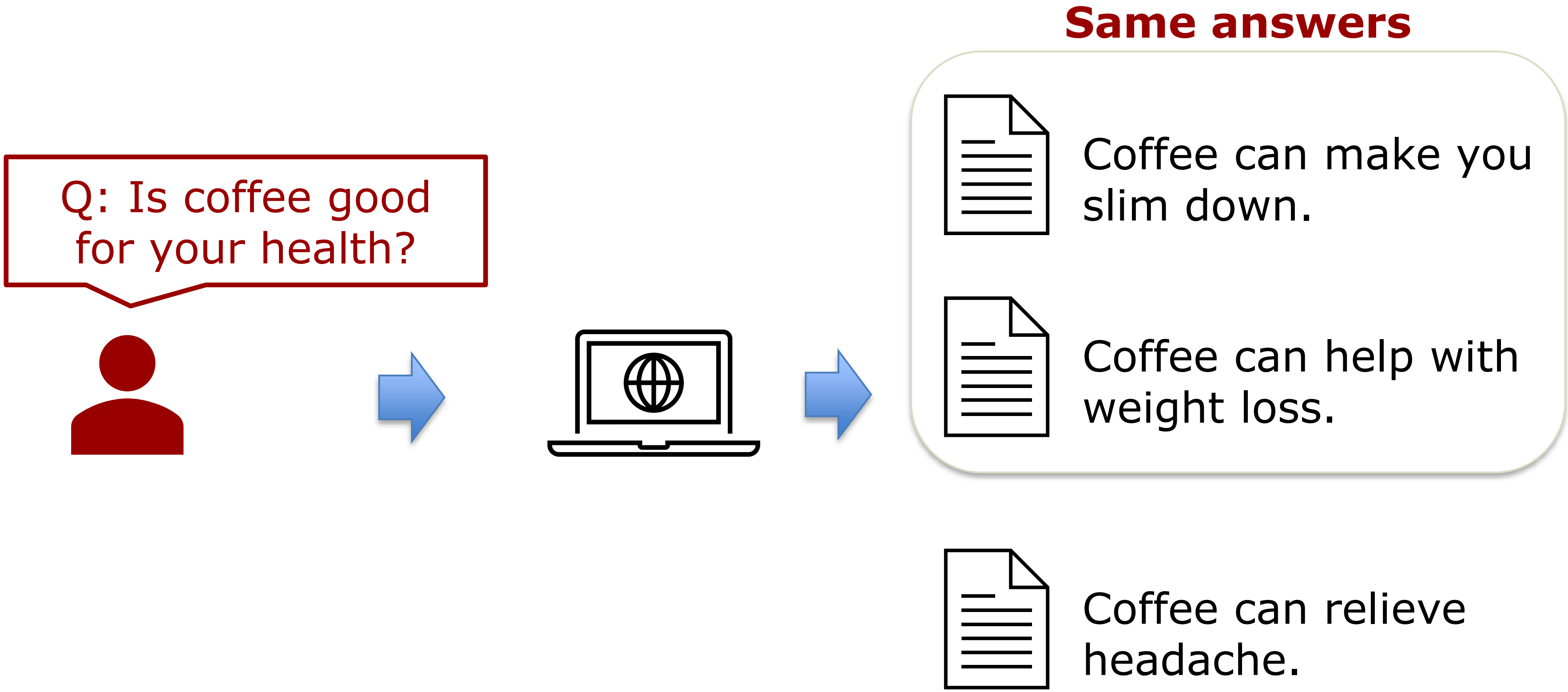
Answer Consolidation: Formulation and Benchmarking

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

Multiple Answers Problem in QA



Q: Is coffee good for your health?

Same answers

Coffee can make you slim down.

Coffee can help with weight loss.

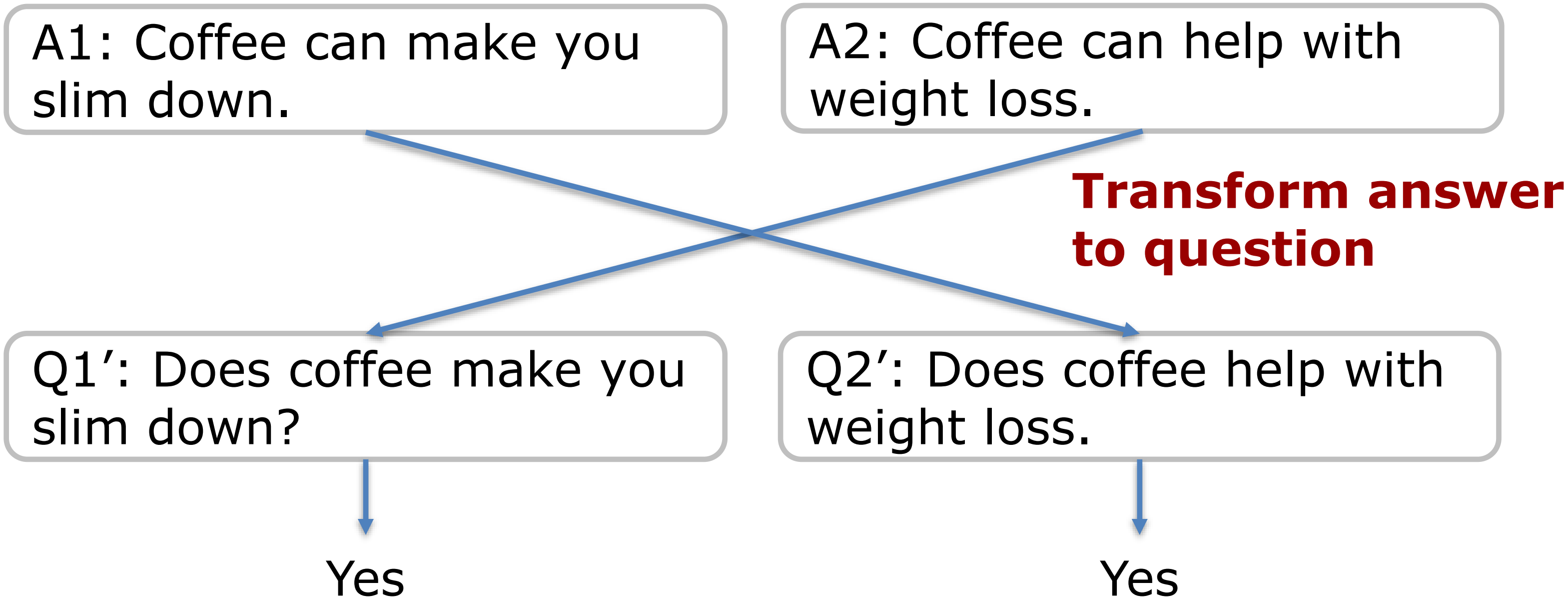
Coffee can relieve headache.

Goal: identify equivalent/distinct answers in QA.

Problem Formulation

Define equivalent/different answers.

Q: Is coffee good for your health?



A1: Coffee can make you slim down.

A2: Coffee can help with weight loss.

Transform answer to question

Q1′: Does coffee make you slim down?

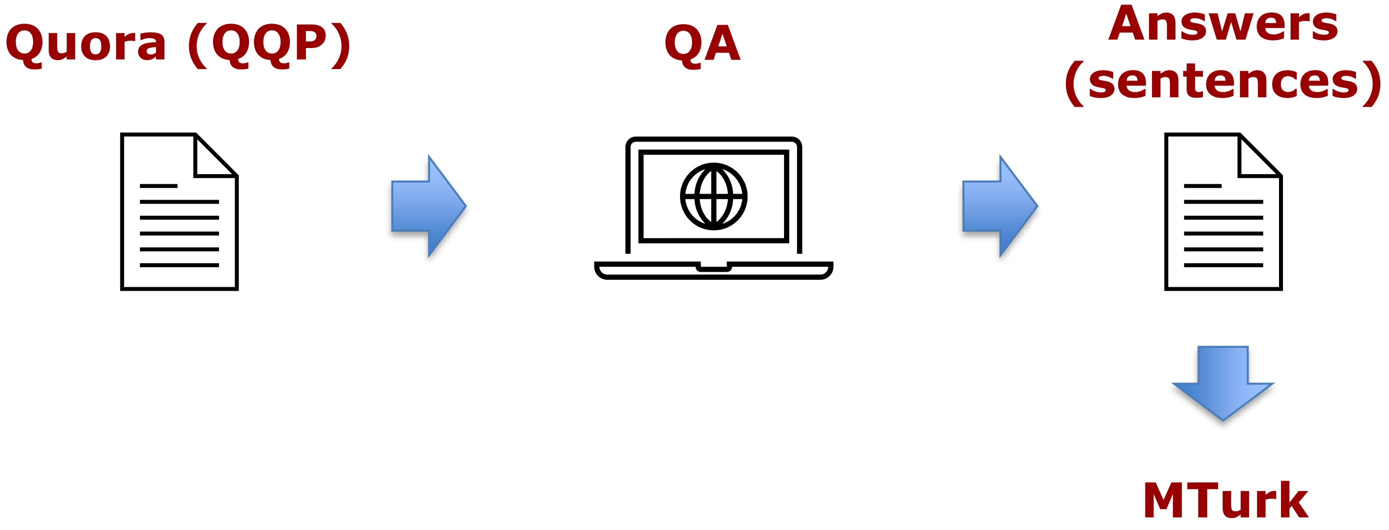
Q2′: Does coffee help with weight loss.

Yes

Yes

Equivalent if answers are **both yes or no**

QUASI Dataset: Construction

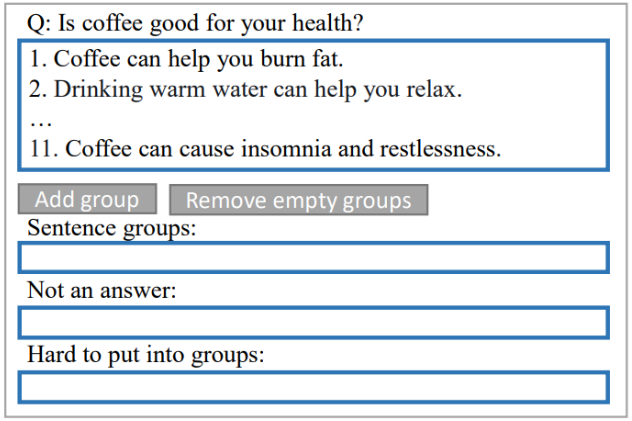


Quora (QQP)

QA

Answers (sentences)

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.

Sentence groups:

Not an answer:

Hard to put into groups:

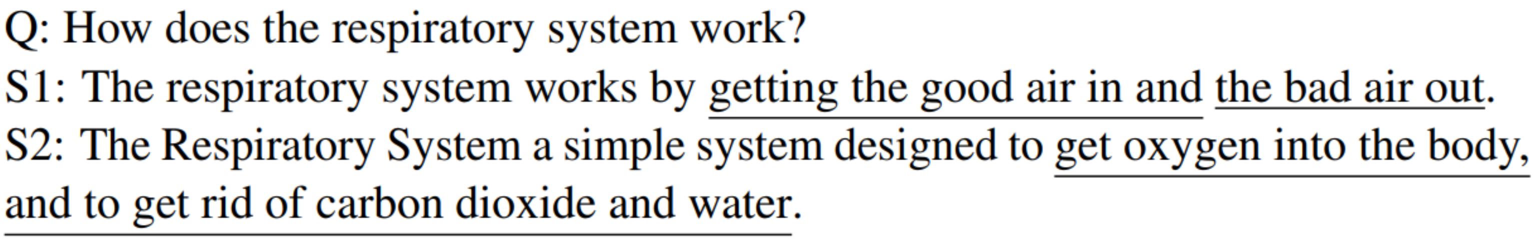
Groups of equivalent answers

QUASI Dataset: Statistics

4,699 questions, 24,006 sentences, and 19,676 groups.

Types of equivalent answers:

1. Formatting + exact match (53%):
   - The answer spans are the same.
2. Lexical variation (11%):
   - The answers spans differ in articles, verb tenses, …
3. Semantic variation (30%):
   - The answer spans have the same semantic meaning, may need external knowledge in identification
   - Example:



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

Settings

1. **Sentence pair classification**
   - Given a question and two answers, decide whether they are in the same group.
2. **Sentence grouping**
   - Put answers into groups.
   - Method: cluster the sentences using the distance of sentence pairs.

Models

Sentence embedding models:

- Inputs:

  <s>$X_q\ X_s$</s>

- Prediction: cosine similarity

Cross-encoders

- Inputs:

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

- Prediction: linear classifier

Answer-aware cross-encoders

- Inputs: extract the answers and add to inputs

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

- Prediction: linear classifier

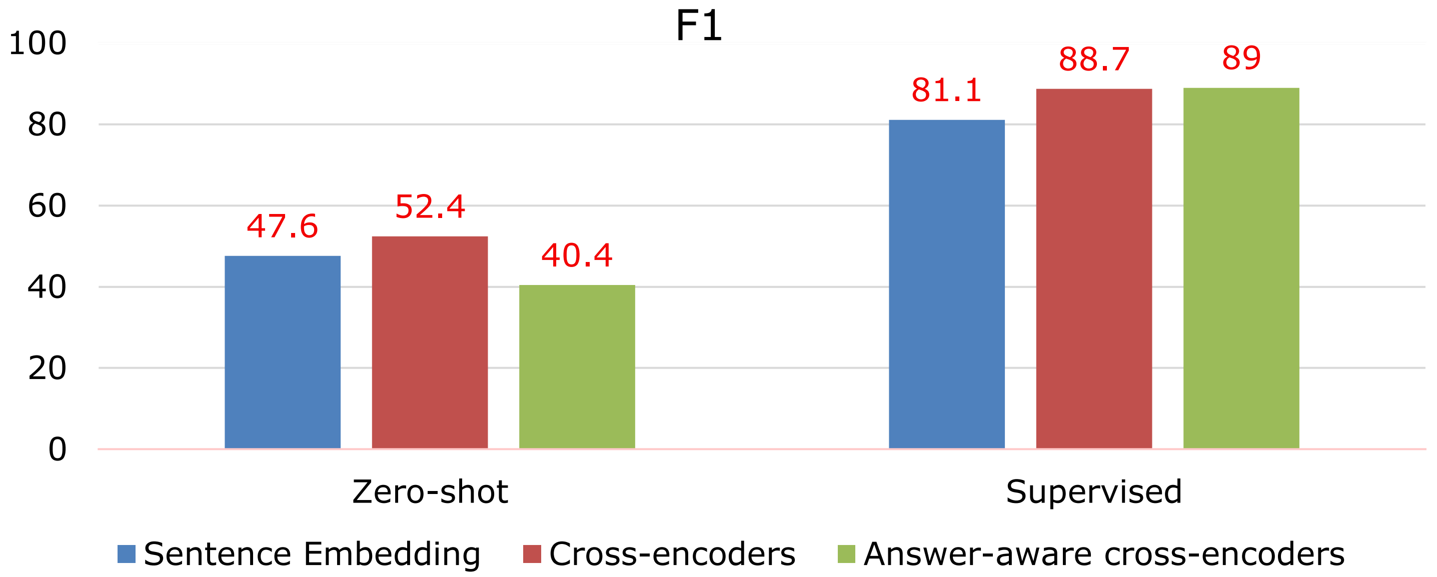
X_q: question

X_s: sentence

X_a: extracted answer

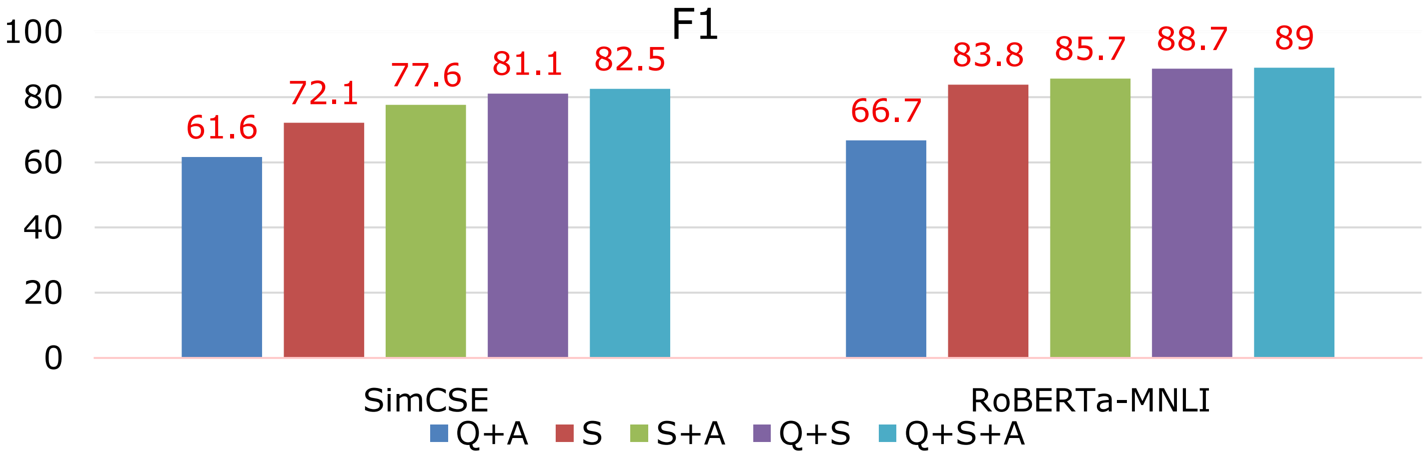
Experiments: Main results

Encoders: SimCSE for sentence embedding, RoBERTa-MNLI for cross-encoders.



Experiments: Ablation

Q: question
S: sentences containing answers
A: answers extracted by UnifiedQA



Conclusion

1. We formulate and propose the **answer consolidation task** that seeks to group answers into equivalent groups.
2. We contribute the **Question-Answer consolidation dataset** (QUASI) for this task and evaluate various models, including sentence embedding models, cross-encoders, and answer-aware cross-encoders.
3. Experiments suggest room for further studies on more **robust and generalizable solutions** for answer consolidation that would largely benefit real-world open-domain QA systems.