facebook AI Research

Unsupervised Question Decomposition for Question Answering

Ethan Perez, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, Douwe Kiela

NYU | ML² Machine Learning for Language

Unsupervised Question Decomposition

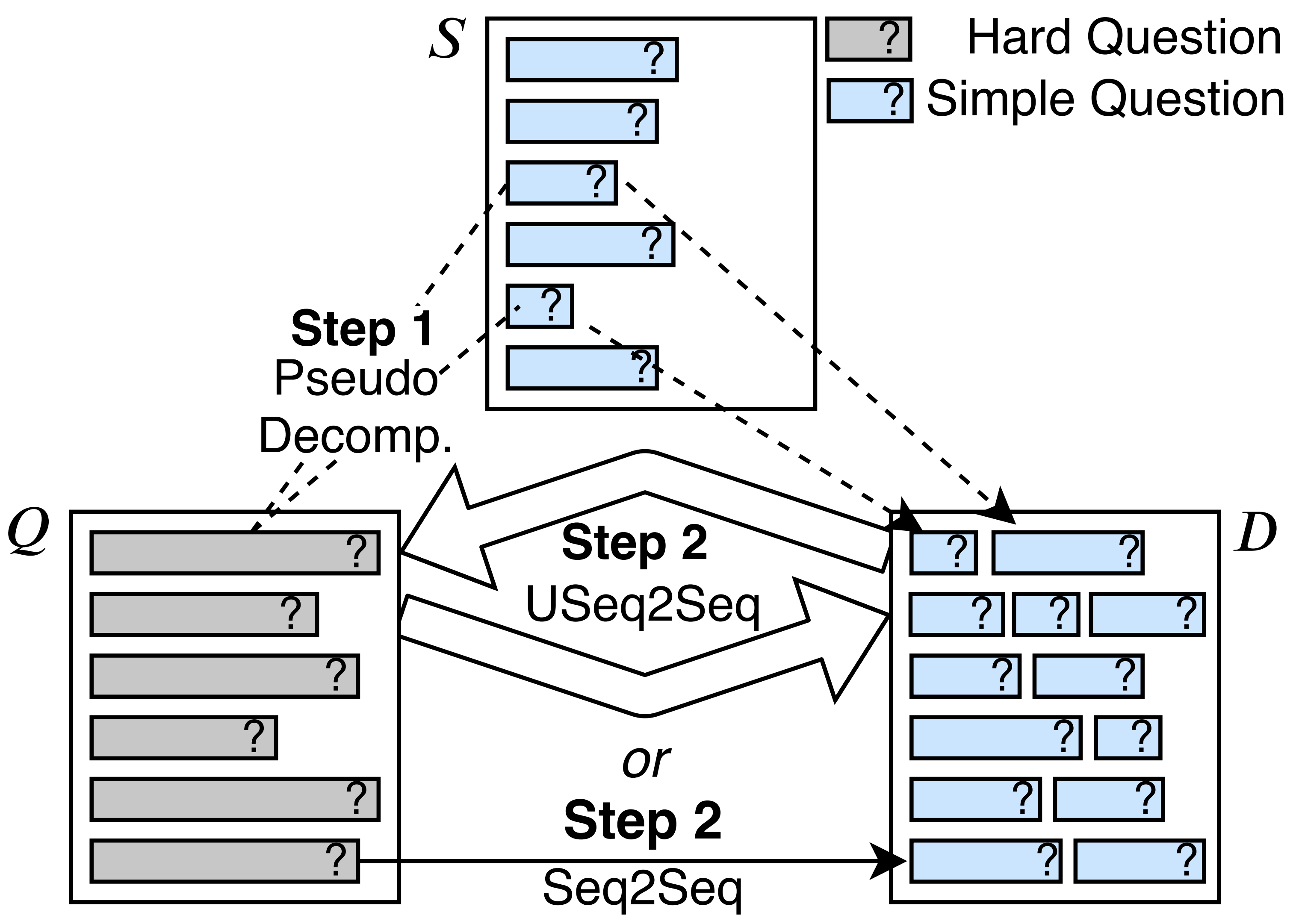


Figure 1

Goal: Improve QA by decomposing hard questions into easier sub-questions that existing QA systems can answer.
Problem: Prior work learns to decompose questions by relying human annotation and extractive heuristics.
Solution: Decompose questions with *unsupervised methods*, using 2 stages (**Figure 1**):

(1) Construct a noisy, "pseudo-decomposition" for each hard question by retrieving relevant sub-question candidates.
(2) Train neural text generation models on that data with standard or unsupervised sequence-to-sequence learning.

Finding: We greatly improve multi-hop QA on HotpotQA with unsupervised decompositions, using a 3-stage method (**Figure 2**):

(1) Generate single-hop sub-questions for a multi-hop question.
(2) Answer sub-questions with a single-hop QA model.
(3) Add sub-questions and their answers as additional input for a multi-hop QA model.

Unsupervised Decomposition

Creating Pseudo-Decompositions

For each **q** in a corpus **Q** of hard questions, we construct a pseudo-decomposition **d'** = [**s**$_1$; **s**$_2$; ... **s**$_N$] by retrieving **s** from a corpus **S** of simple questions. We want **s** that are (1) similar to **q** w.r.t. a metric **f** (e.g., cos distance) and (2) maximally diverse:

$$d'^* = \operatorname*{argmax}_{d' \subset S} \sum_{s_i \in d'} f(q, s_i) - \sum_{\substack{s_i, s_j \in d' \\ s_i \neq s_j}} f(s_i, s_j)$$

We embed **q** and **s** via sum-of-FastText word vectors. We also test random pseudo-decompositions where **s**$_i$ ~ **S**.

Training Models on Pseudo-Decompositions

We train models on pseudo-decompositions via:

- **No Learning**: Use **d'** = [**s**$_1$; **s**$_2$; ... **s**$_N$] as sub-questions
- **Seq2Seq**: maximize *log p(d'|q)*
- **Unsup. Seq2Seq**: learn a **q** → **d** mapping without training on noisy (q, d') pairs, similar to unsupervised translation

QA Results

We greatly improve the baseline by adding sub-questions and answers.

We are competitive with DecompRC, SAE, and HGN which use strong supervision.

Decomp. Method	Pseudo-Decomps.	HotpotQA F1		
		Orig	MultiHop	OOD
✗	✗ (1hop)	66.7	63.7	66.5
✗	✗ (Baseline)	77.0±.2	65.2±.2	67.1±.5
No Learn	Random	78.4±.2	70.9±.2	70.7±.4
	FastText	78.9±.2	72.4±.1	72.0±.1
Seq2Seq	Random	77.7±.2	69.4±.3	70.0±.7
	FastText	78.9±.2	73.1±.2	73.0±.3
USeq2Seq	Random	79.8±.1	76.0±.2	76.5±.2
	FastText	**80.1**±.2	**76.2**±.1	**77.1**±.1
DecompRC*		79.8±.2	76.3±.4	77.7±.2
SAE (Tu et al., 2020) †		80.2	61.1	62.6
HGN (Fang et al., 2019) †		82.2	78.9‡	76.1‡

	Ours	SAE†	HGN†
Test (EM/F1)	66.33/79.34	66.92/79.62	69.22/82.19

Using Decompositions in Question Answering (QA)

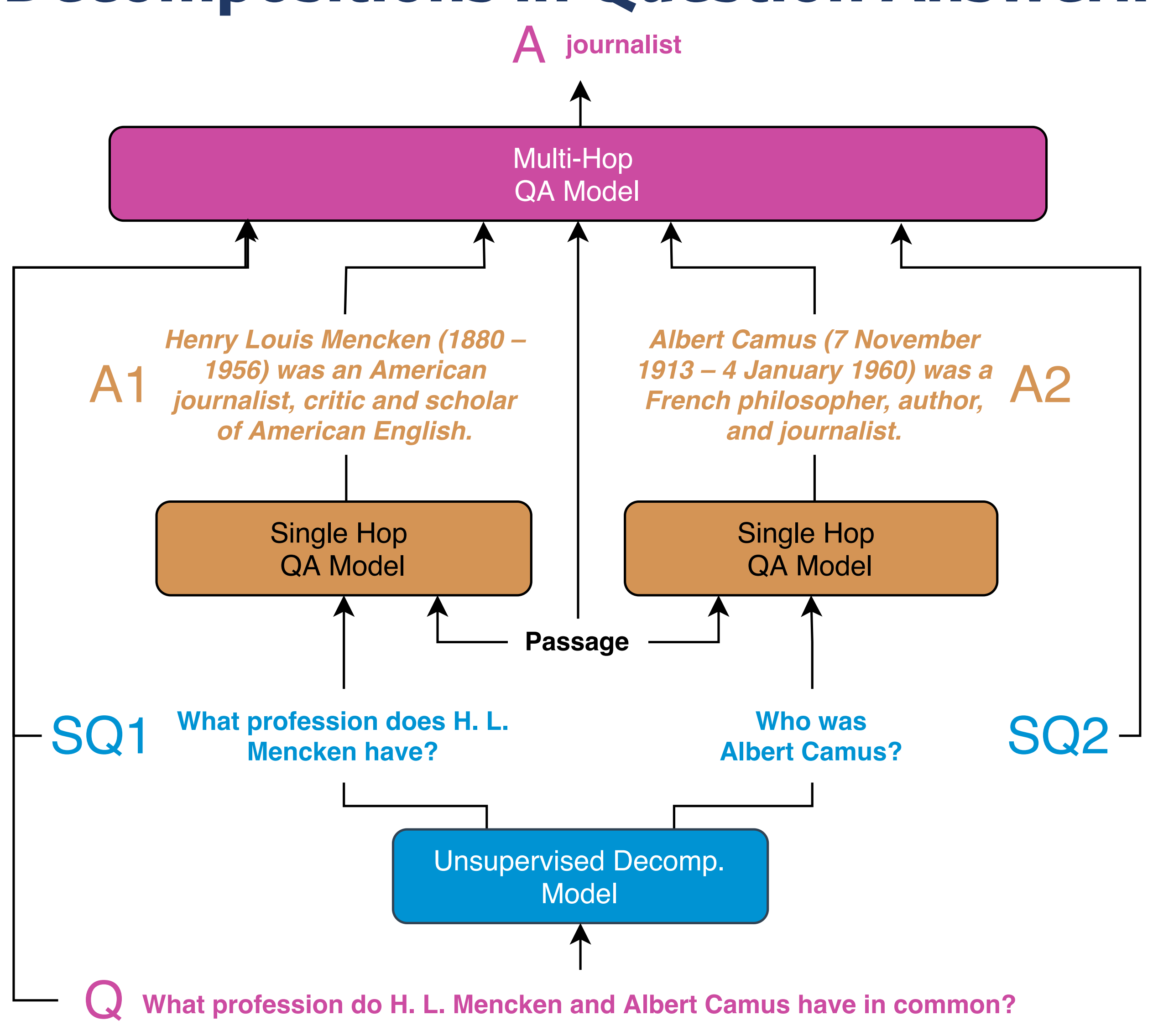

A journalist
Multi-Hop QA Model
A1
Henry Louis Mencken (1880 – 1956) was an American journalist, critic and scholar of American English.
Albert Camus (7 November 1913 – 4 January 1960) was a French philosopher, author, and journalist.
A2
Single Hop QA Model
Single Hop QA Model
Passage
SQ1
What profession does H. L. Mencken have?
Who was Albert Camus?
SQ2
Unsupervised Decomp. Model
Q What profession do H. L. Mencken and Albert Camus have in common?

Figure 2

Examples

Generated sub-questions are single-hop and question-relevant.

Add interpretability to black-box QA models.

Automatically learned to decompose many kinds of questions. Improved QA across all 4 question categories (**Table 1**).

Sub-questions are fluent, especially w.r.t. supervised decompositions (**Table 2**).

Q1: Are both Coldplay and Pierre Bouvier from the same country?
SQ$_1$: Where are Coldplay and Coldplay from?
⌊ Coldplay are a British rock band formed in 1996 by lead vocalist and keyboardist Chris Martin and lead guitarist Jonny Buckland at University College London (UCL).
SQ$_2$: What country is Pierre Bouvier from?
⌊ Pierre Charles Bouvier (born 9 May 1979) is a Canadian singer, songwriter, musician, composer and actor who is best known as the lead singer and guitarist of the rock band Simple Plan.
Â: No

Q2: How many copies of Roald Dahl's variation on a popular anecdote sold?
SQ$_1$: How many copies of Roald Dahl's?
⌊ His books have sold more than 250 million copies worldwide.
SQ$_2$ What is the name of the variation on a popular anecdote?
⌊ "Mrs. Bixby and the Colonel's Coat" is a short story by Roald Dahl that first appeared in the 1959 issue of Nugget.
Â: more than 250 million

Q3: Who is older, Annie Morton or Terry Richardson?
SQ$_1$: Who is Annie Morton?
⌊ Annie Morton (born October 8, 1970) is an American model born in Pennsylvania.
SQ$_2$: When was Terry Richardson born?
⌊ Kenton Terry Richardson (born 26 July 1999) is an English professional footballer who plays as a defender for League Two side Hartlepool United.
Â: Annie Morton

Decomps.	Bridge	Comp.	Intersec.	Single-hop
✗	80.7±.2	73.8±.4	78.1±.6	73.8±.6
✓	**82.3**±.4	**80.1**±.3	**81.2**±.4	**76.7**±.6

Table 1: QA F1 with and without Decompositions

Decomp. Method	GPT2 NLL	% Well-Formed
USeq2Seq	5.56	60.9
DecompRC	6.04	32.6

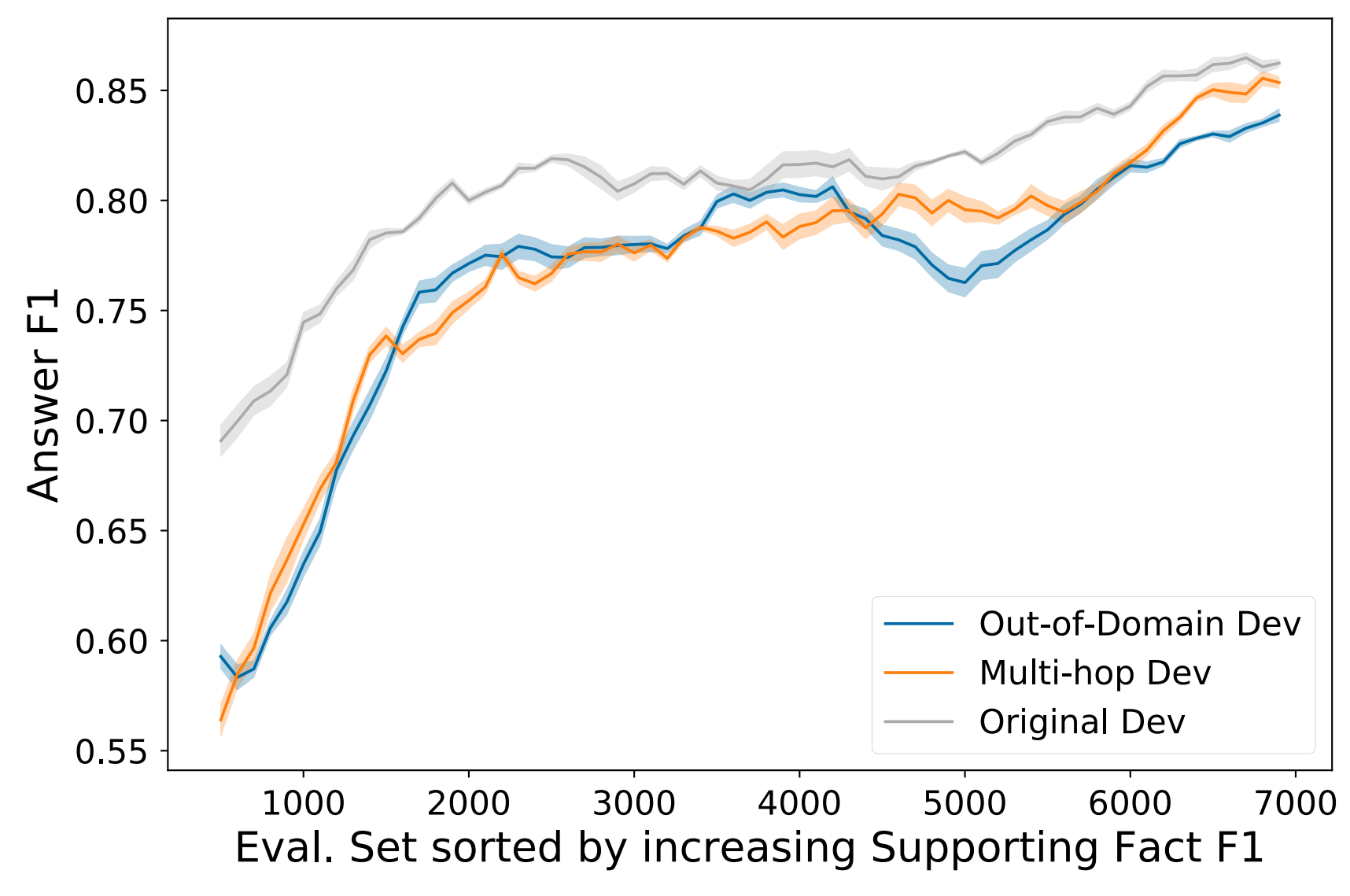
Table 2: Decompositions from USeq2Seq (ours) vs. DecompRC

Analysis

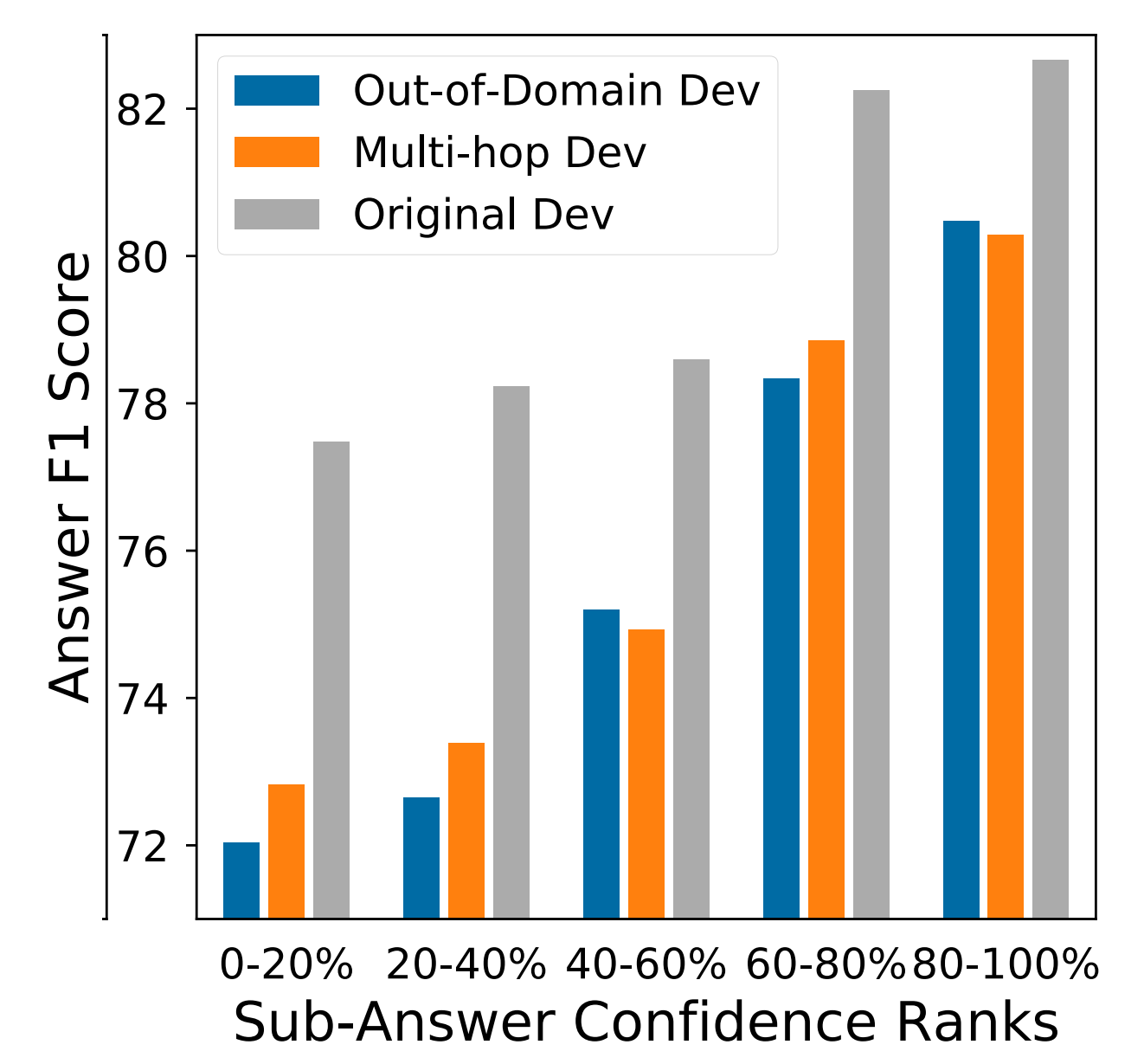
Including sub-answers is crucial. Returning sentences with sub-answer spans is better than just returning sub-answer spans.

SubQs	SubAs	QA F1
✗	✗	77.0±.2
✓	Sentence	80.1±.2
✓	Span	77.8±.3
✓	Random Entity	76.9±.2
✓	✗	76.9±.2
✗	Sentence	80.2±.1

Multi-hop QA improves when the single-hop QA model answers with gold, question-relevant "supporting fact" sentences. We find supporting facts without strong supervision.



Multi-hop QA improves when the single-hop QA model is more confident of its answers to sub-questions. Low confidence sub-answers may be more likely to be incorrect/hurt multi-hop QA.


Out-of-Domain Dev
Multi-hop Dev
Original Dev
Answer F1 Score
Sub-Answer Confidence Ranks