Learning What is Essential in Questions

Daniel Khashabi, Dan Roth (University of Pennsylvania)
Tushar Khot, Ashish Sabharwal (Allen Institute for Artificial Intelligence)

Overview

Challenge: Many QA systems are unable to reliably identify which question words are redundant, irrelevant, or even intentionally distracting. (example) - - - - - →

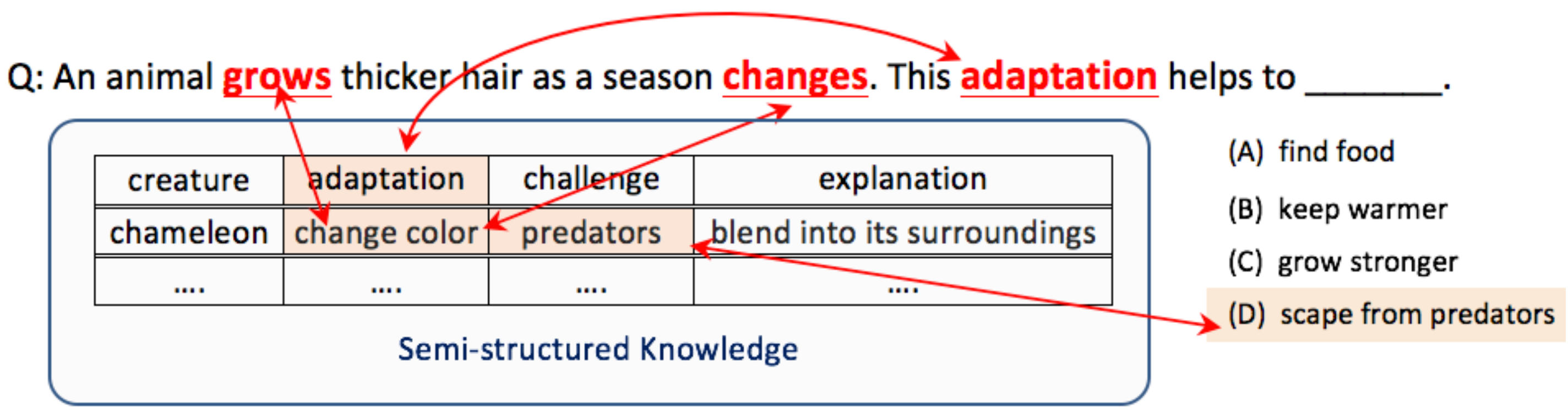
Proposal: Introduce and study the notion of *essential question terms* with the goal of improving such QA systems.

Results: A new ET classifier that reliably (90% mean average precision, MAP) identifies and ranks essential terms in questions, and improves state-of-the-art QA solvers for elementary-level science questions by up to 5%.

1

An animal grows thicker hair as a season changes. This adaptation helps to _______. (A) find food (B) keep warmer (C) grow stronger (D) scape from predators

TableILP (Khashabi et al., 2016) system performs reasoning by aligning the question to semi-structured knowledge. On this question, it aligns only 'grow', 'changes', 'adaptation', resulting in choosing an incorrect answer.



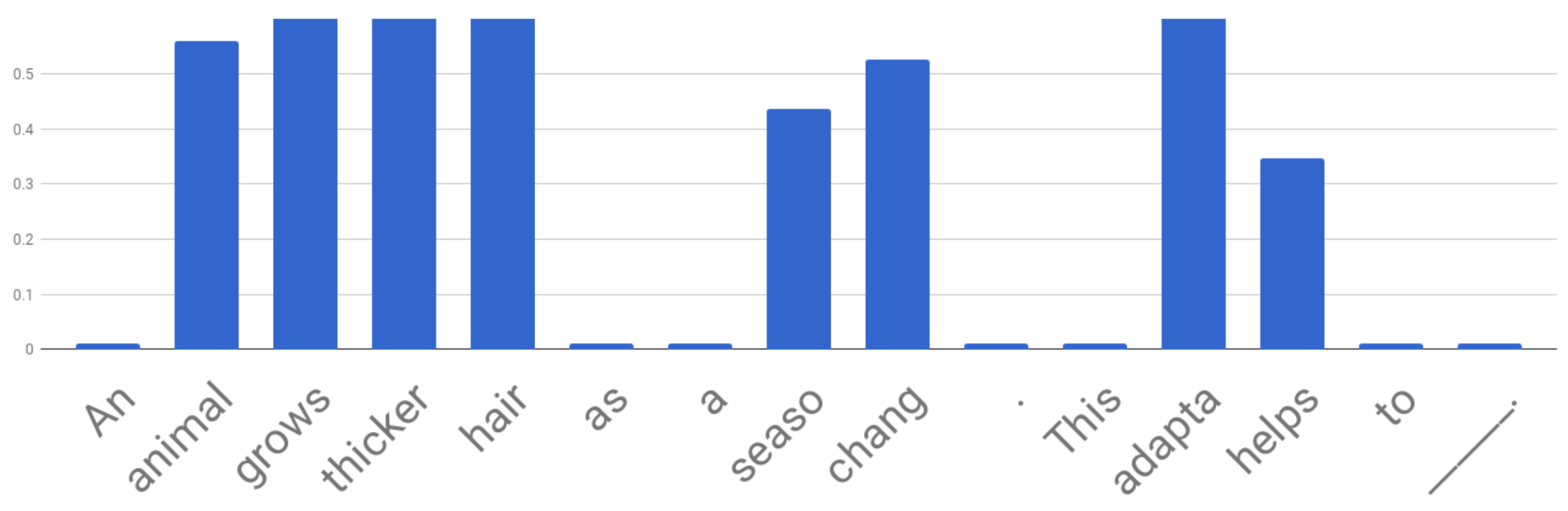
Q: An animal grows thicker hair as a season changes. This adaptation helps to _______.

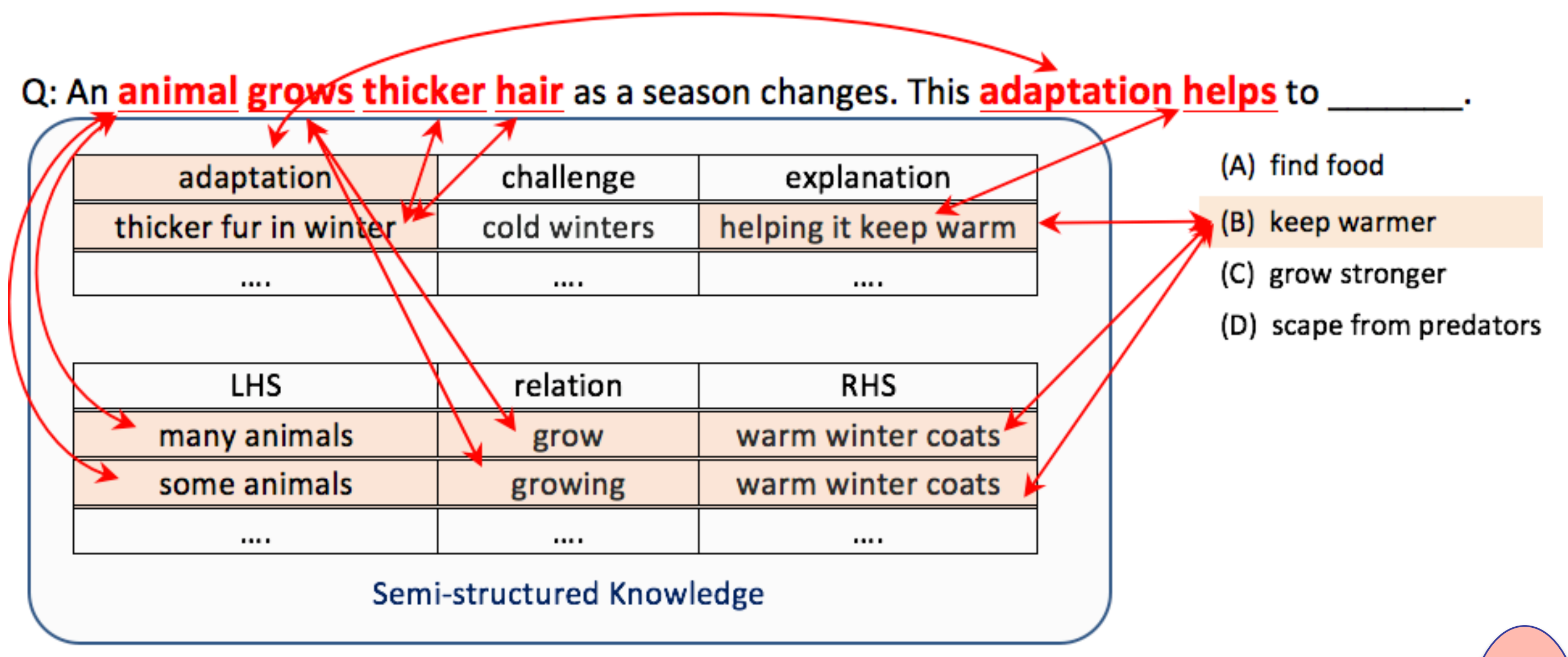
creature	adaptation	challenge	explanation
chameleon	change color	predators	blend into its surroundings
….	….	….	….

Semi-structured Knowledge

(A) find food
(B) keep warmer
(C) grow stronger
(D) scape from predators

Issue: TableILP does not recognize that "thicker hair" is an essential aspect of the question. **Solution**: augment TableILP with essentiality scores:





Q: An animal grows thicker hair as a season changes. This adaptation helps to _______.

adaptation	challenge	explanation
thicker fur in winter	cold winters	helping it keep warm
….	….	….

LHS	relation	RHS
many animals	grow	warm winter coats
some animals	growing	warm winter coats
….	….	….

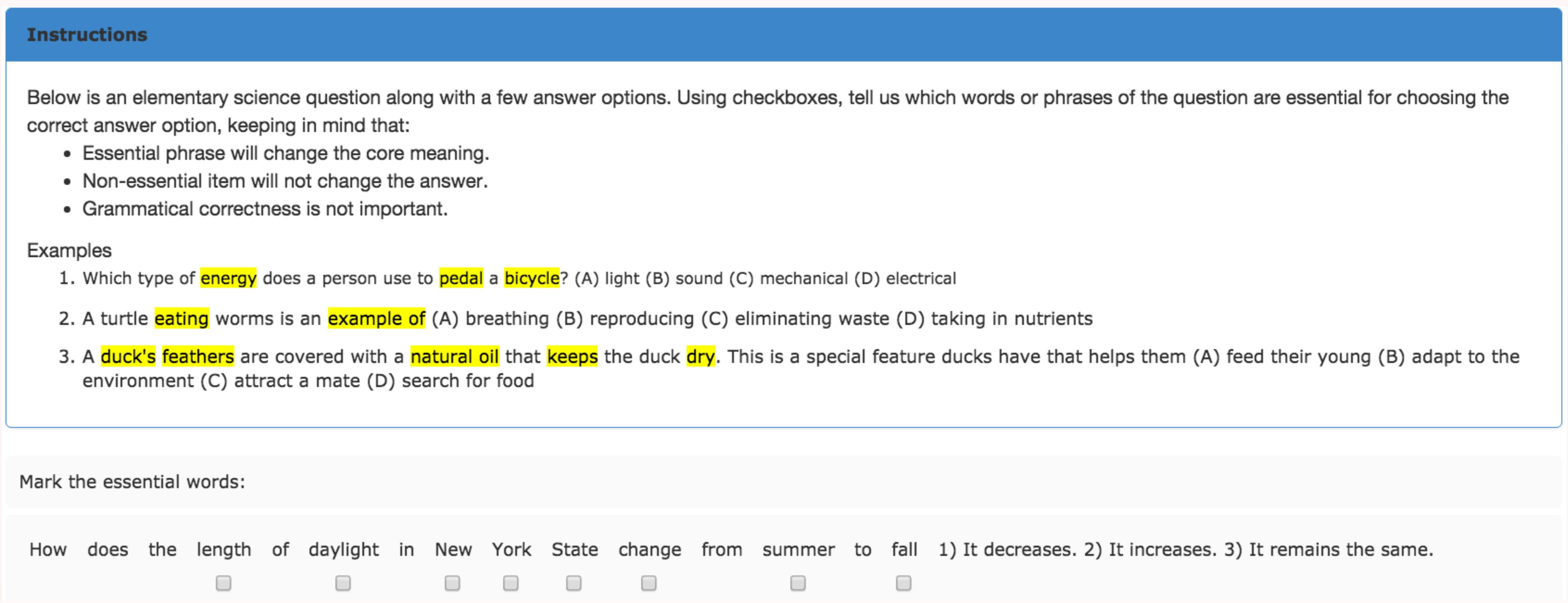
Semi-structured Knowledge

(A) find food
(B) keep warmer
(C) grow stronger
(D) scape from predators

2

Crowd-Sourced Essentiality Dataset

- Collected 2,223 elementary school science exam questions for the annotation.
- The questions were annotated by 5 crowd workers and resulted in 19,380 annotated terms.
- The Fleiss' kappa of κ = 0.58 (inter-annotator agreement very close to 'substantial')



Instructions

Below is an elementary science question along with a few answer options. Using checkboxes, tell us which words or phrases of the question are essential for choosing the correct answer option, keeping in mind that:

- Essential phrase will change the core meaning.
- Non-essential item will not change the answer.
- Grammatical correctness is not important.

Examples

1. Which type of energy does a person use to pedal a bicycle? (A) light (B) sound (C) mechanical (D) electrical
2. A turtle eating worms is an example of (A) breathing (B) reproducing (C) eliminating waste (D) taking in nutrients
3. A duck's feathers are covered with a natural oil that keeps the duck dry. This is a special feature ducks have that helps them (A) feed their young (B) adapt to the environment (C) attract a mate (D) search for food

Mark the essential words:

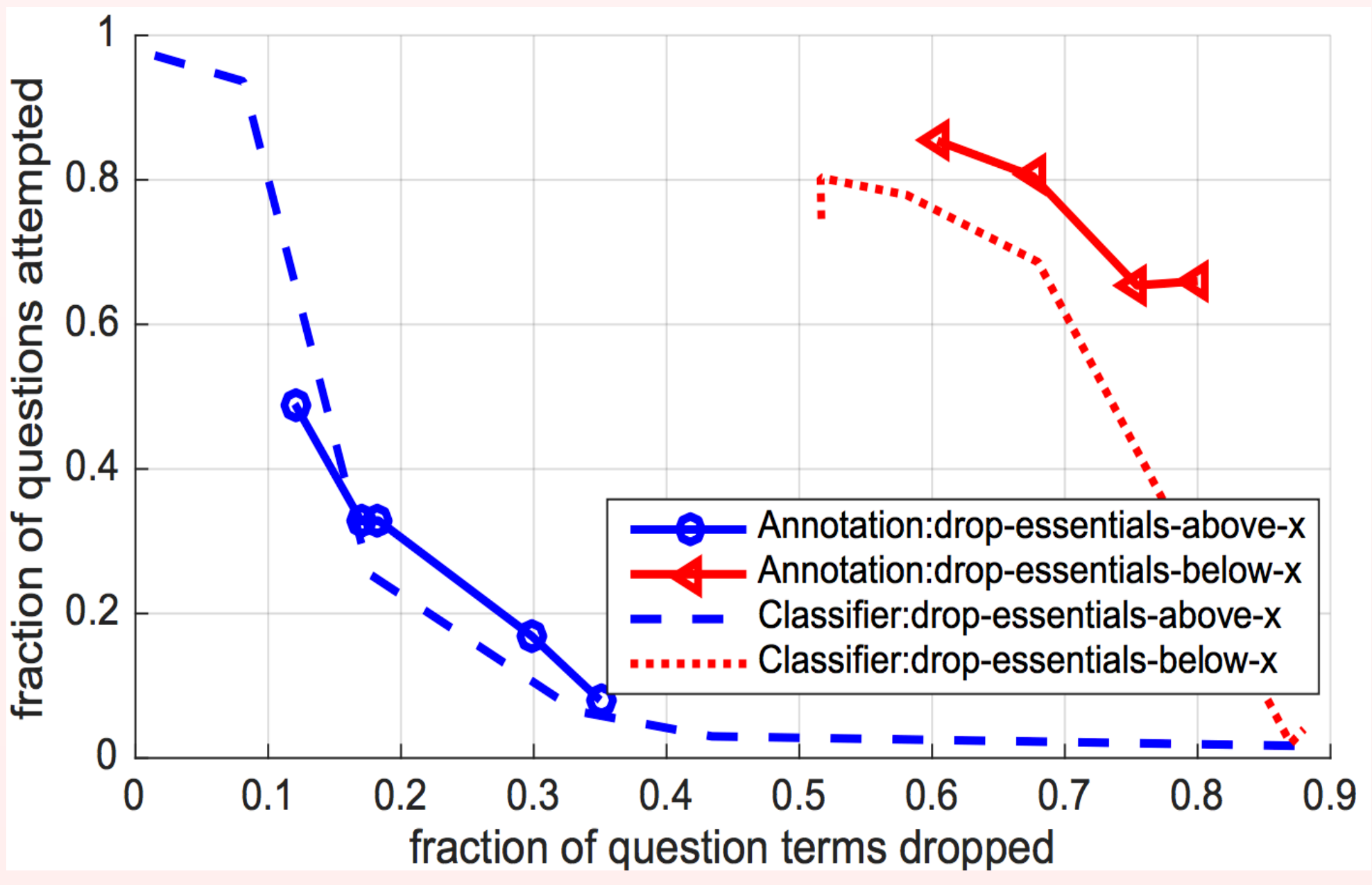
How does the length ☐ of daylight ☐ in New ☐ York ☐ State ☐ change ☐ from summer ☐ to fall ☐ 1) It decreases. 2) It increases. 3) It remains the same.

The Importance of Essential Terms

A crowd-sourcing experiment to validates our hypothesis:

Is the question still *answerable* by a *human,* if a fraction of the essential question terms are *eliminated*?

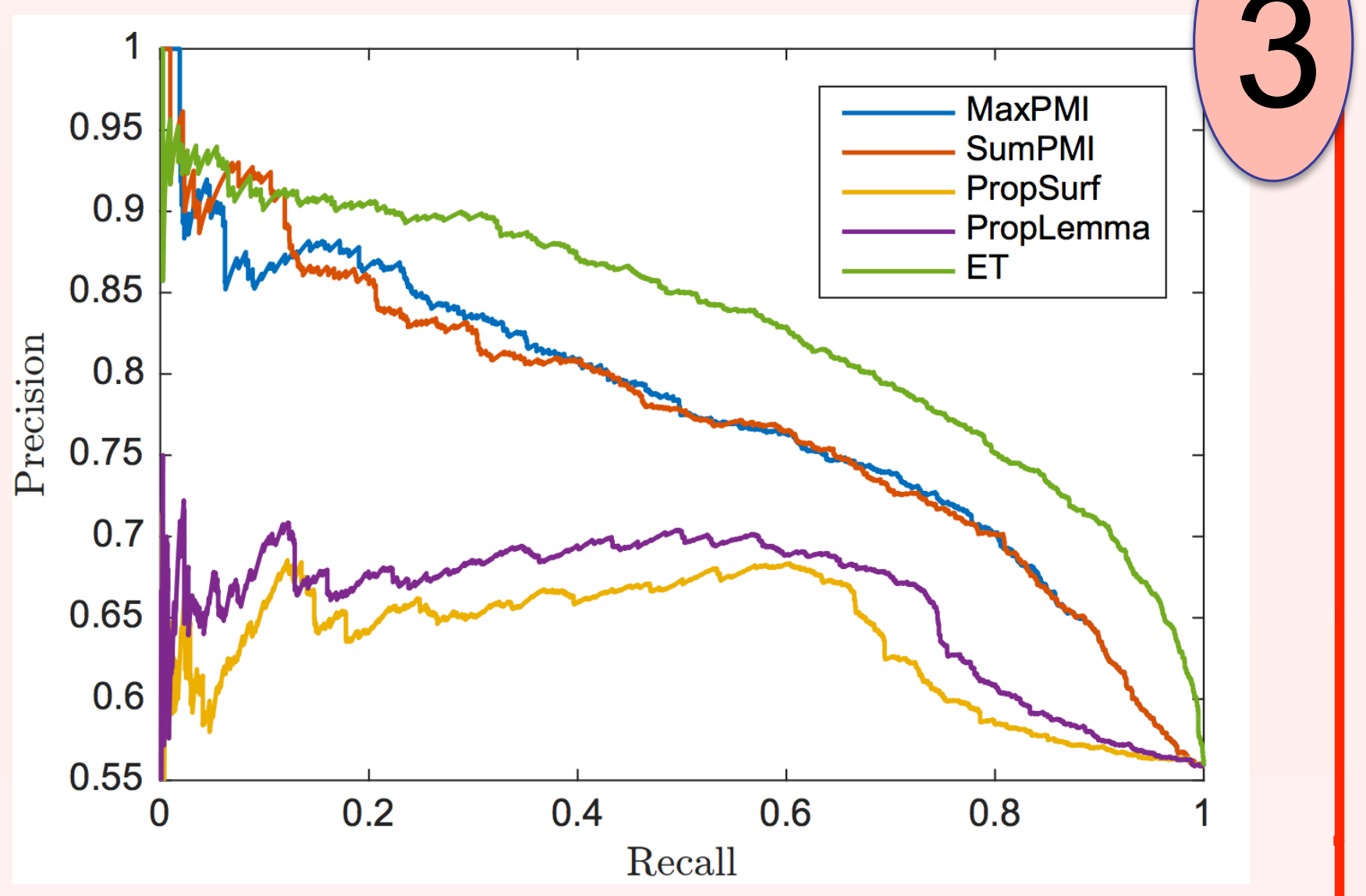
- Average fraction of terms dropped on the horizontal axis and the corresponding fraction of questions attempted on the vertical axis.
- **Blue lines:** effect of eliminating essential sets
- **Red lines:** effect of eliminating non-essentials



Finding #1: **Solid blue line**: dropping even a small fraction of question terms marked as essential dramatically reduces the QA performance of humans.
Finding #2: **Solid red line**: The opposite trend for terms marked as not-essential: even after dropping 80% of such terms, 65% of the questions remained answerable.

3

PR curves for methods as the threshold is varied



ET Classifier has 5% higher AUC (area under the curve) and outperforms baselines by ~5% across the precision-recall spectrum.

Learning Essential Terms

ET Classifier

Given a question q, answer options a, we seek a classifier that predicts whether a given term is essential.

- Trained a linear SVM on real-valued essentiality scores binarized to 1 if they are at least 0.5, and to 0 otherwise.
- Features include syntactic (e.g., dependency parse based) and semantic (e.g., Brown cluster representation of words, a list of scientific words) properties of question words, as well as their combinations. In total, we use 120 types of features.

Baselines

- **Supervised:** Score for a term is proportional to times it was marked as essential in the annotated dataset. *PropSurf:* based on surface string; *PropLem:* based on lemmatizing surface string.
- **Unsupervised:** *MaxPMI* and *SumPMI* score the importance of a word x by max-ing or summing, resp., PMI scores *p(x, y)* across all answer options *y* for *q*.

Binary Classification of Terms.

- Consider all question terms pooled together, resulting in a dataset of 19,380 terms annotated independently as essential or not.
- Evaluate binary predictions on these terms.

ET classifier achieves an F1 score of 0.80, which is 5%-14% higher than the baselines. Its accuracy at 0.75 is statistically significantly better than all baselines based on the Binomial exact test (p-value 0.05).

System	F1
MaxPMI	0.75
SumPMI	0.75
PropSurf	0.66
PropLem	0.69
ET Classifier	**0.80**

Ranking Question Terms.

- Evaluate how well systems rank all question terms in order of essentiality.
- For the ranked list produced by each classifier for each question, we compute the average precision, and take the mean of these AP values across questions to obtain the mean average precision (MAP) score.

ET classifier achieves a MAP of 90.2%, which is 3%-5% higher than baselines, and demonstrates that one can learn to reliably identify essential question terms.

System	MAP
MaxPMI	0.87
SumPMI	0.85
PropSurf	0.85
PropLem	0.86
ET Classifier	**0.90**

End-to-end QA

4

IR Solver + ET

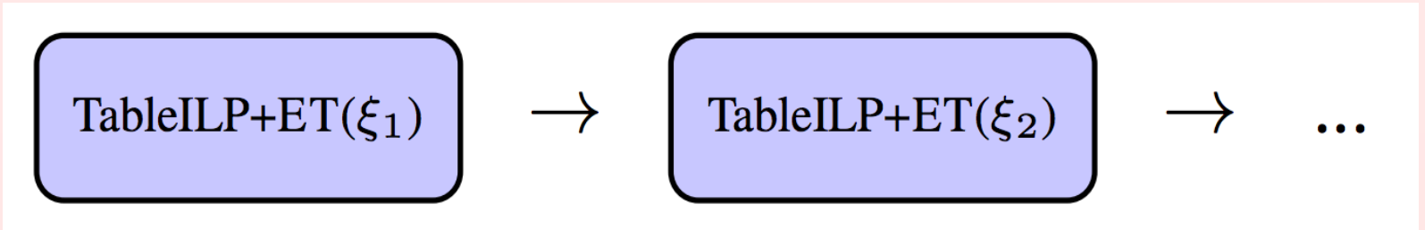
A modified IR system where it query (q', a), with q' being the *essential* subset of q.

Dataset	Basic IR	IR+ET
Regents	59.11	**60.85**
AI2Public	57.90	**59.10**
RegtsPertd	61.84	**66.84**

On RegtsPertd set, ET improves IR by 4.26% to 63.4%. Previous state-of-the-art solver, TableILP, achieves a score of 61.5%, thus achieving a new state of the art.

TableILP + ET

We employ a cascade system that starts with a strong essentiality requirement and progressively weakens it:



Questions unanswered by the first system are delegated to the second, and so on.

On a dataset adversarially generated by analyzing TableILP's mistakes, TableILP + ET corrects 41.7% of the mistakes.

This error-reduction illustrates that the extra attention mechanism added to TableILP via the concept of essential question terms helps it cope with distracting terms.

Authors acknowledge the support from AI2 and DARPA (under the DEFT program).