E-KAR: A Benchmark for Rationalizing Natural Language Analogical Reasoning

Jiangjie Chen[1,2], Rui Xu[1], Ziquan Fu[3], Wei Shi[4], Zhongqiao Li[1], Xinbo Zhang[2], Changzhi Sun[2], Lei Li[5], Yanghua Xiao[1] and Hao Zhou[2]

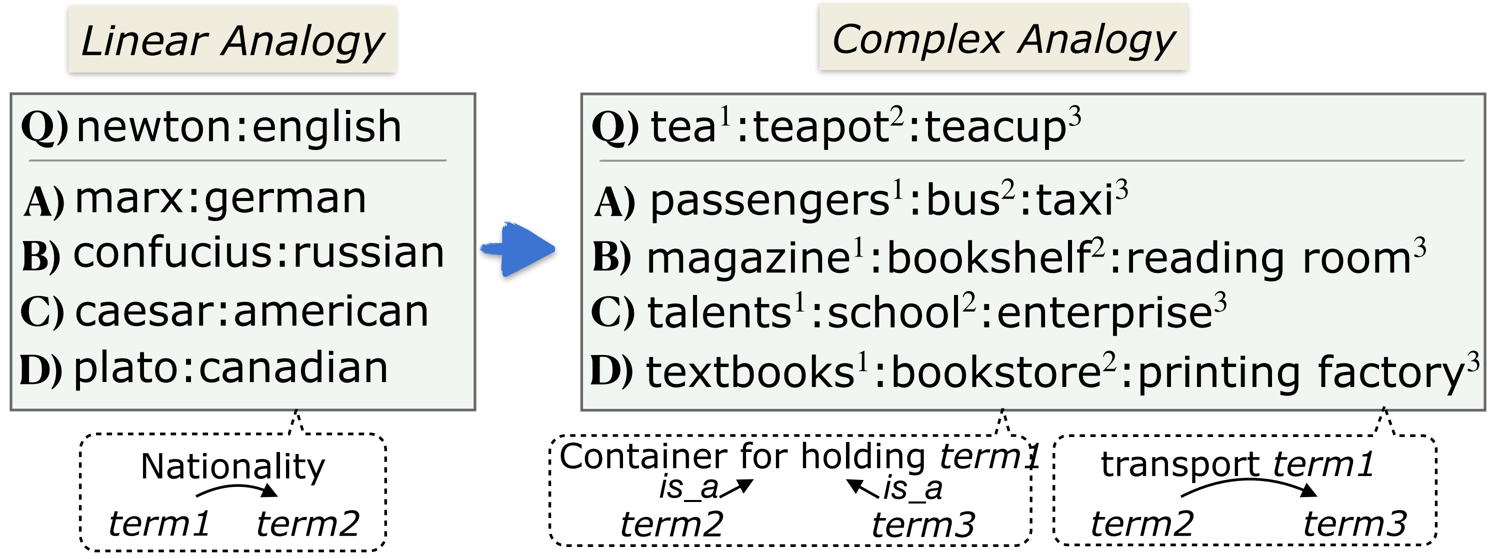
[1]Fudan University [2]ByteDance AI Lab [3]Brain Technologies Inc.
[4]South China University of Technology [5]University of California, Santa Barbara

Project Page

Introduction

Task: Word Analogy Recognition

- *From **linear analogy** to **complex analogy***
- Benchmarking and explaining complex and knowledge-intensive analogical reasoning.



Linear Analogy

Q) newton:english
A) marx:german
B) confucius:russian
C) caesar:american
D) plato:canadian

Complex Analogy

Q) tea^1:teapot^2:teacup^3
A) passengers^1:bus^2:taxi^3
B) magazine^1:bookshelf^2:reading room^3
C) talents^1:school^2:enterprise^3
D) textbooks^1:bookstore^2:printing factory^3

The Limitations of Previous Work

- ***Methods***: Hold a connectionist assumption
  - $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} = \overrightarrow{queen}$
- ***Benchmarks***: Evaluate pre-trained word representations for linear analogy
  - Binary Relations: Lexical, morphological, semantic.
  - Not explainable

The Motivations of This Work

- 🏆 *for Reasoning: Being Right for the Right Reasons*
- Rationalize reasoning with rationales that reveal the analogical reasoning process
- 🤔 *Human-like analogical reasoning requires human-level analogical benchmarks*

Contributions

*We propose a novel benchmark **E-KAR** (**E**xplainable **K**nowledge-intensive **A**nalogical **R**easoning) for rationalizing natural language analogical reasoning, which is:*

- ***Challenging***: E-KAR requires intensive commonsense, factual and cultural knowledge to solve, as well as reasoning skills.
- ***Explainable***: E-KAR is *manually annotated* with free-text explanations based on *structure-mapping theory* to justify analogical reasoning.
- ***Bilingual***: E-KAR is in both Chinese and English.

The E-KAR Benchmark

Challenging | *Sourced from Civil Service Exams*

Why are analogical problems from CSE *challenging*?

Knowledge-intensive term relations

1. Linguistic knowledge
2. Commonsense knowledge
3. Encyclopedic/factual knowledge
4. Cultural knowledge
5. Relations of three terms
6. Negated facts

husband:job
- *Husband is **not** a job.*

car:tires
- *A car is **not** made of tires.*
- *A car consists of tires.*

Explainable | *Manual Free-text Explanations*

How to rationalize analogical reasoning?

Structure-mapping theory *(Minnameier et al, 2010)*

Abduction → **Mapping** → **Validation**

Verbalize the process into free-text.

How to design and acquire the rationales?

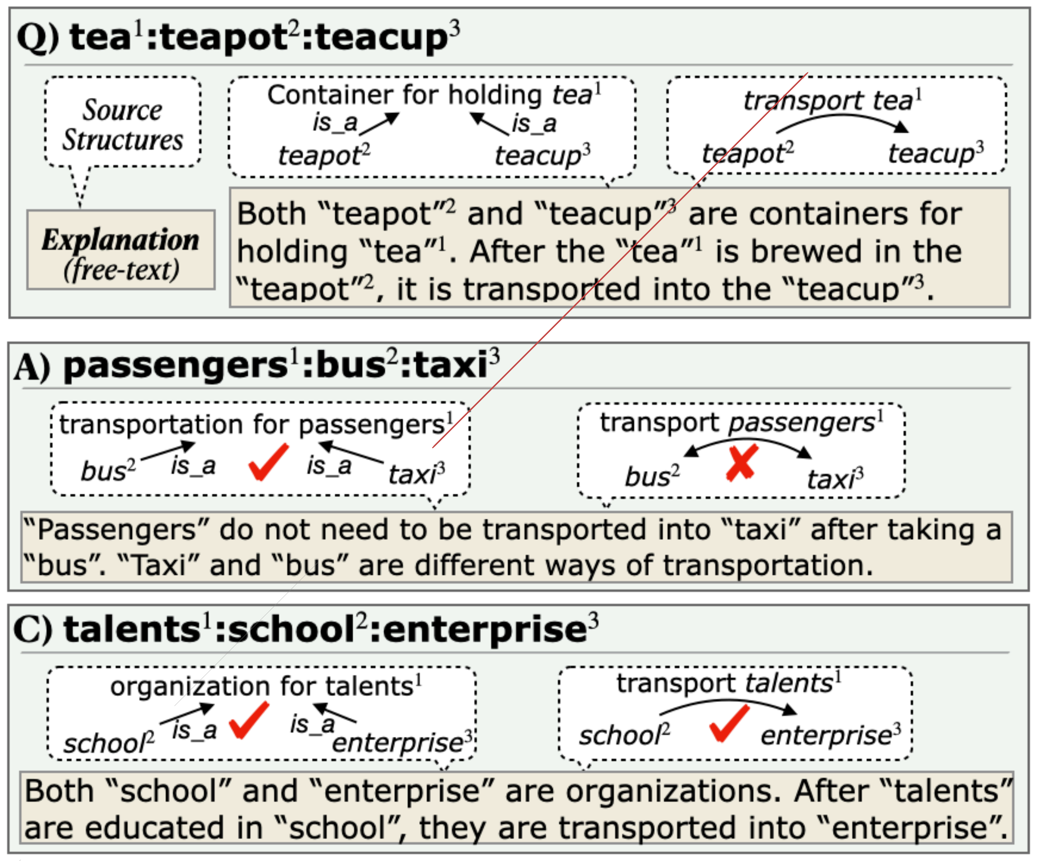
Human-annotated Free-text

Double-checking Strategy for quality control

Explanation for Every Query and Candidate

Both *Refuting* & *Supporting* Explanation

With Evidence Showing Why



Q) tea^1:teapot^2:teacup^3

Both "teapot"^{2} and "teacup"^{3} are containers for holding "tea"^{1}. After the "tea"^{1} is brewed in the "teapot"^{2}, it is transported into the "teacup"^{3}.

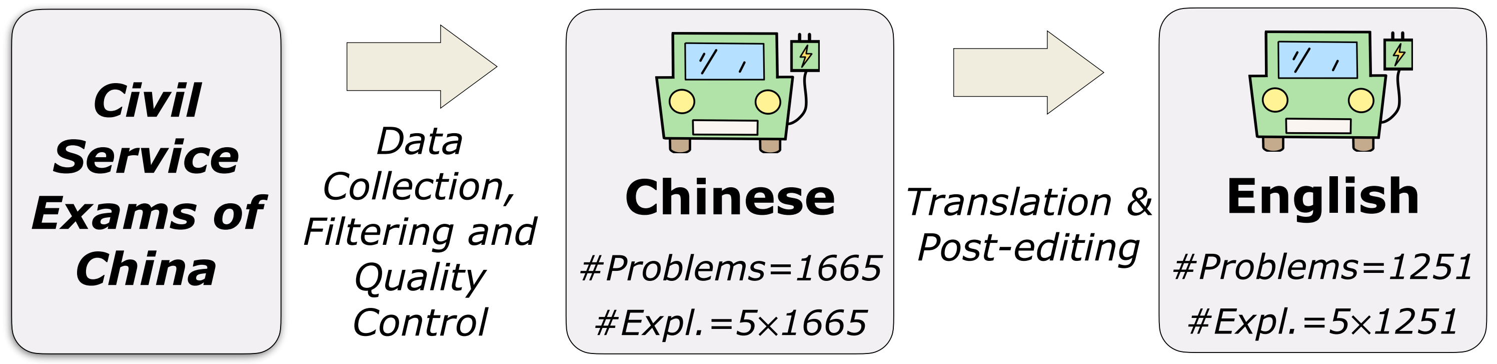
A) passengers^1:bus^2:taxi^3

"Passengers" do not need to be transported into "taxi" after taking a "bus". "Taxi" and "bus" are different ways of transportation.

C) talents^1:school^2:enterprise^3

Both "school" and "enterprise" are organizations. After "talents" are educated in "school", they are transported into "enterprise".

Bilingual | *Chinese & English*



Civil Service Exams of China

Data Collection, Filtering and Quality Control

Chinese
#Problems=1665
#Expl.=5×1665

Translation & Post-editing

English
#Problems=1251
#Expl.=5×1251

Preliminary Explorations

Two Shared Tasks

Task 1: Analogical Question Answering

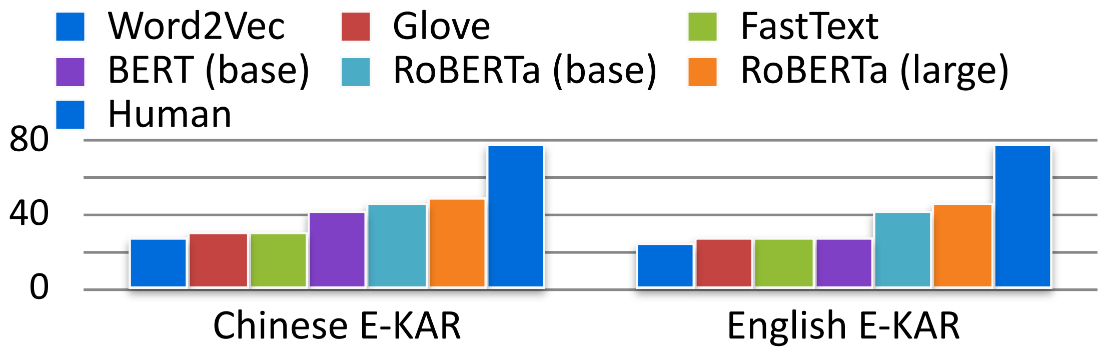
- **Task type**: multiple-choice question answering
- **Input**: Query + Candidates
- **Output**: Correct Choice
- **Evaluation**:
  - *QA Acc.*

Task 2: Explanation Generation

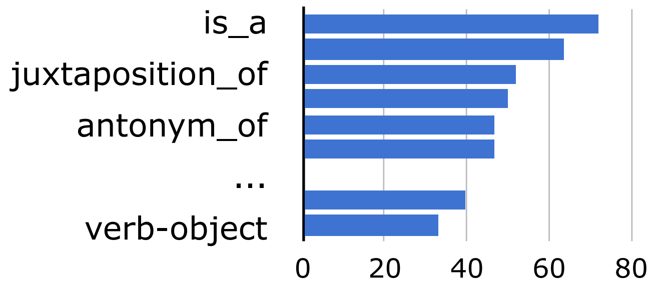
- **Task type**: text generation
- **Input**: Query + Candidates
- **Output**: Free-text explanations for both query $\mathscr{E}_Q$ and candidates $\mathscr{E}_A$
- **Evaluation**:
  - ~~ROUGE, BERTScore, …~~ (unreliable)
  - *Rationalized QA Acc.* (Acc. with $\mathscr{E}$)

Better metrics for explanations needed !

***Lesson 1**: W2Vs and LMs both struggle at complex analogical reasoning.*

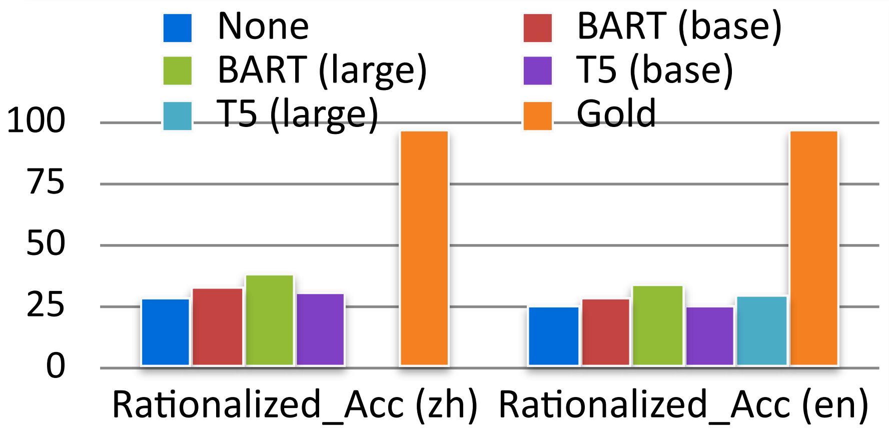


Humans outperform SOTA models by large margins.



Most QA errors occur on semantic relations, which demands heavily on commonsense and factual knowledge and reasoning skills.

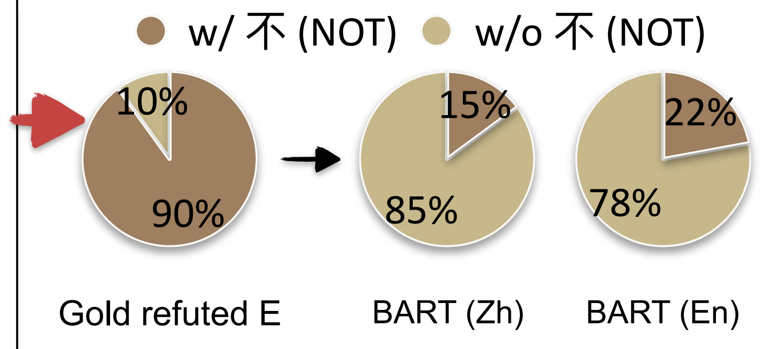
***Lesson 2**: Generative LMs struggle at rationalizing analogical reasoning.*



1. *Poor quality of generated explanations, improvement over baseline but fall far behind gold.*
2. *Gold explanations can be exploited by Analogical QA models to achieve nearly perfect results (97.7%).*

Error Analysis

1. *Unable to generate negated facts for refutation.*
2. *Generating factually incorrect statements.*
3. *Biasing towards common patterns.*



If you have any questions, please email: jjchen19@fudan.edu.cn