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

Jiangjie Chen, Rui Xu, Ziquan Fu, Wei Shi, Zhongqiao Li, Xinbo Zhang, Changzhi Sun, Lei Li, Yanghua Xiao, Hao Zhou
Word Analogy Recognition

<table>
<thead>
<tr>
<th>Query</th>
<th>Candidate answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q) newton:english</td>
<td>A) marx:german</td>
</tr>
<tr>
<td></td>
<td>B) confucius:russian</td>
</tr>
<tr>
<td></td>
<td>C) caesar:american</td>
</tr>
<tr>
<td></td>
<td>D) plato:canadian</td>
</tr>
</tbody>
</table>

An analogical reasoning problem from The Bigger Analogy Test Set (BATS).
Word Analogy Recognition

Word analogy as multiple-choice QA

<table>
<thead>
<tr>
<th>Query</th>
<th>Candidate answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q) newton:english</td>
<td>A) marx:german</td>
</tr>
<tr>
<td></td>
<td>B) confucius:russian</td>
</tr>
<tr>
<td></td>
<td>C) caesar:american</td>
</tr>
<tr>
<td></td>
<td>D) plato:canadian</td>
</tr>
</tbody>
</table>

An analogical reasoning problem from The Bigger Analogy Test Set (BATS).
From Linear Analogy to Complex Analogical Reasoning

\[ \text{king} - \text{man} + \text{woman} = \text{queen} \]

Linear Analogy (Ethayarajh et al. 2019)

e.g. Word2Vec
From Linear Analogy to Complex Analogical Reasoning

**Linear Analogy**  
(Ethayarajh et al. 2019)

\[
\vec{king} - \vec{man} + \vec{woman} = \vec{queen}
\]

*Previous work*

- **Methods:** Hold a connectionist assumption

E.g. Word2Vec
From Linear Analogy to Complex Analogical Reasoning

**Linear Analogy** (Ethayarajh et al. 2019)

\[ \overrightarrow{\text{king}} - \overrightarrow{\text{man}} + \overrightarrow{\text{woman}} = \overrightarrow{\text{queen}} \]

**Previous work**

- **Methods**: Hold a connectionist assumption
- **Benchmarks**: Evaluate pre-trained word representations for linear analogy

**Simple Binary Relations**

- Lexical, morphological, simple semantic relations.

E.g., Word2Vec

Nationality

term1 → term2
From Linear Analogy to Complex Analogical Reasoning

**Linear Analogy** (Ethayarajh et al. 2019)

\[
\text{king} - \text{man} + \text{woman} = \text{queen}
\]

**Previous work**

- **Methods:** Hold a connectionist assumption
- **Benchmarks:** Evaluate pre-trained word representations for linear analogy

**Simple Binary Relations**

Lexical, morphological, simple semantic relations.

**Not Explainable**

Unable to reveal human-like analogical reasoning process.

e.g. Word2Vec

Nationality

\[ \text{term1} \rightarrow \text{term2} \]
From Linear Analogy to Complex Analogical Reasoning

**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\(^3\)

An analogical reasoning problem from **Civil Service Exams of China**.
(Translated)
The E-KAR Benchmark

**Challenging**
Sourced from Civil Service Exams of China

**Explainable**
Free-text Explanations

**Bilingual**
Chinese & English

---

**Complex Analogy**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Lang.</th>
<th>Data Size (train / val / test)</th>
<th># of Terms in Cand.</th>
<th>Has Expl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAT</td>
<td>En</td>
<td>0 / 37 / 337</td>
<td>2</td>
<td>×</td>
</tr>
<tr>
<td>Google</td>
<td>En</td>
<td>0 / 50 / 500</td>
<td>2</td>
<td>×</td>
</tr>
<tr>
<td>BATS</td>
<td>En</td>
<td>0 / 199 / 1,799</td>
<td>2</td>
<td>×</td>
</tr>
</tbody>
</table>

**E–KAR**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Lang.</th>
<th>Data Size (train / val / test)</th>
<th># of Terms in Cand.</th>
<th>Has Expl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zh</td>
<td>1,155 / 165 / 335</td>
<td>2(64.5%)</td>
<td>✅</td>
<td></td>
</tr>
<tr>
<td>En</td>
<td>870 / 119 / 262</td>
<td>2(60.5%)</td>
<td>✅</td>
<td></td>
</tr>
</tbody>
</table>

---

#Problems=1665
#Expl. =5×1665

#Problems=1251
#Expl. =5×1251
The E-KAR Benchmark

**Challenging**
Sourced from Civil Service Exams of China

**Explainable**
Free-text Explanations

**Bilingual**
Chinese & English

---

**Knowledge-intensive term relations**

1. Linguistic knowledge
2. Commonsense knowledge
3. Encyclopedic/factual knowledge
4. Cultural knowledge
5. Relations of three terms (35% vs. 0%)
6. Negated facts

---

**Commonsense Knowledge**
- transport tea
  - teapot
  - teacup

**Factual Knowledge**
- cause
  - Lunar & Solar gravity
  - Tide

**Chinese Idioms:**
- 路见不平：拔刀相助
  - [Translation] One the road is rough: draw a knife to help
  - [Meaning] See someone is in trouble: do one’s best to help

---

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

- **car:tires**
  - A car is not made of tires.
  - A car consists of tires.
The E-KAR Benchmark

Challenging
Sourced from Civil Service Exams of China

Explainable
Free-text Explanations

Bilingual
Chinese & English

🤔 How to Rationalize Analogical Reasoning?
Analogical Reasoning: A Psychological Perspective

Structure-mapping theory
(Minnameier et al, 2010)

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$^3$
Analogical Reasoning: A Psychological Perspective

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

**Abduction**
Draw a *source structure* that may work for target.

**Q)** tea¹:teapot²:teacup³

<table>
<thead>
<tr>
<th>Source Structures</th>
<th>Container for holding tea¹</th>
<th>transport tea¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>is_a</td>
<td>is_a</td>
<td></td>
</tr>
<tr>
<td>teapot²</td>
<td>teacup³</td>
<td></td>
</tr>
</tbody>
</table>

**A)** passengers¹:bus²:taxi³

**B)** magazine¹:bookshelf²:reading room³

**C)** talents¹:school²:enterprise³

**D)** textbooks¹:bookstore²:printing factory³
Analogical Reasoning: A Psychological Perspective

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

**Abduction**
- Draw a *source structure* that may work for target.

**Mapping**
- Map the structure to the target domain.

### Q) tea<sup>1</sup>:teapot<sup>2</sup>:teacup<sup>3</sup>
- Container for holding tea<sup>1</sup>
- transport tea<sup>1</sup>

### A) passengers<sup>1</sup>:bus<sup>2</sup>:taxi<sup>3</sup>
- transportation for passengers<sup>1</sup>
- transport passengers<sup>1</sup>

### B) magazine<sup>1</sup>:bookshelf<sup>2</sup>:reading room<sup>3</sup>
- ?

### C) talents<sup>1</sup>:school<sup>2</sup>:enterprise<sup>3</sup>
- organization for talents<sup>1</sup>
- transport talents<sup>1</sup>

### D) textbooks<sup>1</sup>:bookstore<sup>2</sup>:printing factory<sup>3</sup>
- organization
- transport textbooks<sup>1</sup>
Analogical Reasoning: A Psychological Perspective

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

Abduction
Draw a source structure that may work for target.

Mapping
Map the structure to the target domain.

Validation
Validity check and justification w.r.t. solving the target problem.

**Q) tea\(^1\): teapot\(^2\): teacup\(^3\)**

- **Source Structures**
  - Container for holding tea\(^1\)
  - is_a
  - teapot\(^2\)
  - is_a
  - teacup\(^3\)
  - transport tea\(^1\)

- **Transport Structures**
  - teapot\(^2\)
  - teacup\(^3\)

**A) passengers\(^1\): bus\(^2\): taxi\(^3\)**

- transportation for passengers\(^1\)
- bus\(^2\)
- is_a
- taxi\(^3\)
- transport passengers\(^1\)
- bus\(^2\)
- x
- taxi\(^3\)

**B) magazine\(^1\): bookshelf\(^2\): reading room\(^3\)**

- bookshelf\(^2\)
- x
- reading room\(^3\)

**C) talents\(^1\): school\(^2\): enterprise\(^3\)**

- organization for talents\(^1\)
- school\(^2\)
- is_a
- enterprise\(^3\)
- transport talents\(^1\)
- school\(^2\)
- ✓
- enterprise\(^3\)

**D) textbooks\(^1\): bookstore\(^2\): printing factory\(^3\)**

- organization
- bookstore\(^2\)
- ✓
- printing factory\(^3\)
- transport textbooks\(^1\)
- bookstore\(^2\)
- x
- printing factory\(^3\)
How to Rationalize Analogical Reasoning?

🏆 for Reasoning: Being Right for the Right Reasons

**Abduction**

**Mapping**

**Validation**

The Right Reasons: Verbalize the structure-mapping process into free-text explanations.

<table>
<thead>
<tr>
<th>Q) $\text{tea}^1: \text{teapot}^2: \text{teacup}^3$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source Structures</strong></td>
</tr>
<tr>
<td><strong>Container for holding tea</strong></td>
</tr>
<tr>
<td>$\text{is}_a$</td>
</tr>
<tr>
<td>$\text{teapot}^2$</td>
</tr>
<tr>
<td>$\text{teacup}^3$</td>
</tr>
<tr>
<td><strong>transport tea</strong></td>
</tr>
<tr>
<td>$\text{teapot}^2$</td>
</tr>
<tr>
<td>$\text{teacup}^3$</td>
</tr>
<tr>
<td><strong>Explanation (free-text)</strong></td>
</tr>
<tr>
<td>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”$^3$.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A) $\text{passengers}^1: \text{bus}^2: \text{taxi}^3$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>transportation for passengers</strong></td>
</tr>
<tr>
<td>$\text{bus}^2$</td>
</tr>
<tr>
<td>$\text{is}_a$</td>
</tr>
<tr>
<td>$\text{taxi}^3$</td>
</tr>
<tr>
<td><strong>transport passengers</strong></td>
</tr>
<tr>
<td>$\text{bus}^2$</td>
</tr>
<tr>
<td>$\times$</td>
</tr>
<tr>
<td>$\text{taxi}^3$</td>
</tr>
<tr>
<td><strong>“Passengers” do not need to be transported into “taxi” after taking a “bus”. “Taxi” and “bus” are different ways of transportation.</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C) $\text{talents}^1: \text{school}^2: \text{enterprise}^3$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>organization for talents</strong></td>
</tr>
<tr>
<td>$\text{school}^2$</td>
</tr>
<tr>
<td>$\text{is}_a$</td>
</tr>
<tr>
<td>$\text{enterprise}^3$</td>
</tr>
<tr>
<td><strong>transport talents</strong></td>
</tr>
<tr>
<td>$\text{school}^2$</td>
</tr>
<tr>
<td>$\checkmark$</td>
</tr>
<tr>
<td>$\text{enterprise}^3$</td>
</tr>
<tr>
<td><strong>Both “school” and “enterprise” are organizations. After “talents” are educated in “school”, they are transported into “enterprise”.</strong></td>
</tr>
</tbody>
</table>
"Passengers" do not need to be transported into "taxi" after taking a "bus". "Taxi" and "bus" are different ways of transportation.

Example

Refutation

Evidence

Challenging
Sourced from Civil Service Exams of China

Explainable
Free-text Explanations

Bilingual
Chinese & English

Human-annotation
Free-text

Double-checking Strategy for quality control

(Please check the paper for details.)
The E-KAR Benchmark

Challenging
Sourced from Civil Service Exams of China

Explainable
Free-text Explanations

Bilingual
Chinese & English

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

Refutation

Evidence

Human-annotation Free-text

Explanation for Every Query and Candidate

Double-checking Strategy for quality control

Both Refuting & Supporting Explanation

(Please check the paper for details.)
The E-KAR Benchmark

Challenging
Sourced from Civil Service Exams of China

Explainable
Free-text Explanations

Bilingual
Chinese & English

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

Example

Evidence

Human-annotation Free-text

Explanation for Every Query and Candidate
Both Refuting & Supporting Explanation

Double-checking Strategy for quality control

With Evidence Showing Why

(Please check the paper for details.)
The E-KAR Benchmark

**Challenging**
Sourced from Civil Service Exams of China

**Explainable**
Free-text Explanations

**Bilingual**
Chinese & English

---

**Civil Service Exams of China**

Data Collection, Filtering and Quality Control

---

**Chinese**

#Problems=1665
#Expl.=5×1665

**English**

#Problems=1251
#Expl.=5×1251

---

(Please check the paper for details.)
Lesson 1: W2Vs and LMs both struggle at complex analogical reasoning.

Humans outperform SOTA models by large margins.

(Please check the paper for details.)
Lessons from Preliminary Exploration of E-KAR

Lesson 2: LMs struggle at rationalizing analogical reasoning.

Reminder

- **Explanation Generation**
  - **Input**: Query + Candidates
  - **Output**: Free-text explanations for both query $E_Q$ and candidates $E_A$
- **Evaluation**: The performance gain in QA when prompted with generated explanations ($E_Q$ & $E_A$)
  - **Main Metric**: Rationalized QA Accuracy (Acc. with $E$)
  - **QA Model**: RoBERTa-large

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%).*
Lesson 2: LMs struggle at rationalizing analogical reasoning.

Error Analysis

1. Unable to generate negated facts for refutation.
2. Generating factually incorrect statements.

Ex1. "term1" and "term2" has the same meaning.
Ex2. "term1" is a "term2".

(Please check the paper for details.)
What is Next?

• **What we have**: A novel benchmark for rationalizing analogical reasoning, which is *challenging*, *explainable* and *bilingual*.

• Analogical reasoning by effectively interacting with various kinds of knowledge.
  – e.g. commonsense, factual and *cultural* knowledge.

• Generate reasons with evidence to rationalize reasoning.
  – Particularly, enable models to generate *negated* statements/facts.
Have Fun with E-KAR!

https://ekar-leaderboard.github.io

https://eval.ai/web/challenges/challenge-page/1671/overview

jjchen19@fudan.edu.cn

https://jiangjiechen.github.io