# Benchmarking Knowledge-Enhanced Commonsense Question Answering via Knowledge-to-Text Transformation

边宁, 韩先培, 陈波*, 孙乐
中文信息处理实验室
AAAI-2021, * chenbo@iscas.ac.cn, 13552104889

## 1. Introduction

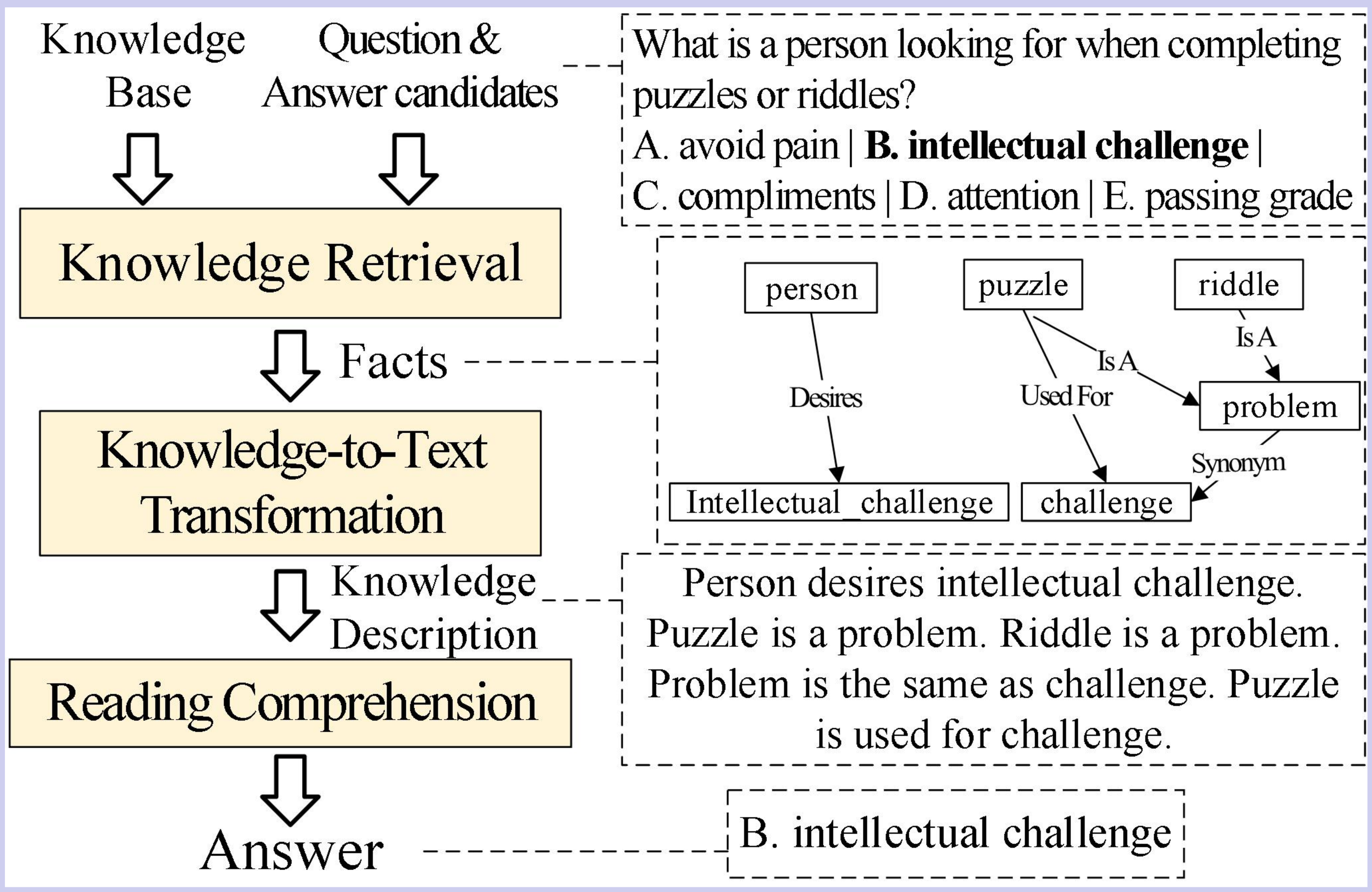
Commonsense Question Answering (**CQA**): Answering questions whose answers rely on commonsense knowledge.

*Where on a river can you hold a cup* ***upright*** *to catch water on a sunny day?*
✔***waterfall***, ✘*bridge*, ✘*valley*, ✘*pebble*, ✘*mountain*

CQA relays on commonsense **knowledge**:
(waterfall, *IsA*, vertical flow of moving water) (upright, *synonym*, vertical)

**We Try to Answer Three Important Questions in CQA**:
Q1: How far can we get by exploiting external knowledge for CQA?
Q2: How much potential of knowledge has been exploited in current CQA models?
Q3: Which are the most promising directions for future CQA?

## 2. Knowledge-to-Text CQA Model




**Three knowledge-to-text transformation methods**

| | |
|---|---|
| Knowledge path | Puzzle → *IsA* → Problem → *Synonym* → Challenge |
| Template-based | Puzzle is a problem. Problem is the same as challenge. |
| Paraphrasing-based | Puzzles are problems. The problem is the same as the challenge. |
| Retrieval-based | Puzzle problem is a challenge game for children. |

## 3. Benchmarking Experiments

| Model | Knowledge Source | BERT | XLNet | RoBERTa | ALBERT |
|---|---|---|---|---|---|
| **Human** | -- | 88.9 | 88.9 | 88.9 | 88.9 |
| **Golden Knowledge** | Human Explanations | 81.1 | 85.1 | 84.7 | 83.7 |
| **Knowledge-to-Text** | | | | | |
| Template-based | ConceptNet | 67.9 | 77.5 | 78.1 | 81.1 |
| Paraphrasing-based | ConceptNet | 67.2 | 74.9 | 77.8 | 79.3 |
| Retrieval-based | ConceptNet | 65.0 | 75.0 | 77.1 | 79.4 |
| Full | ConceptNet | **70.4** | **80.3** | **80.8** | **83.3** |
| **Best Knowledge-enhanced System with Different PLMs** | ConceptNet | 69.0 (Ma et al. 2019) | 79.3 (Lv et al. 2019) | 80.8 (KEDGN) | (No available model so far) |
| **Base Model** | No knowledge | 63.6 | 68.9 | 76.2 | 78.6 |

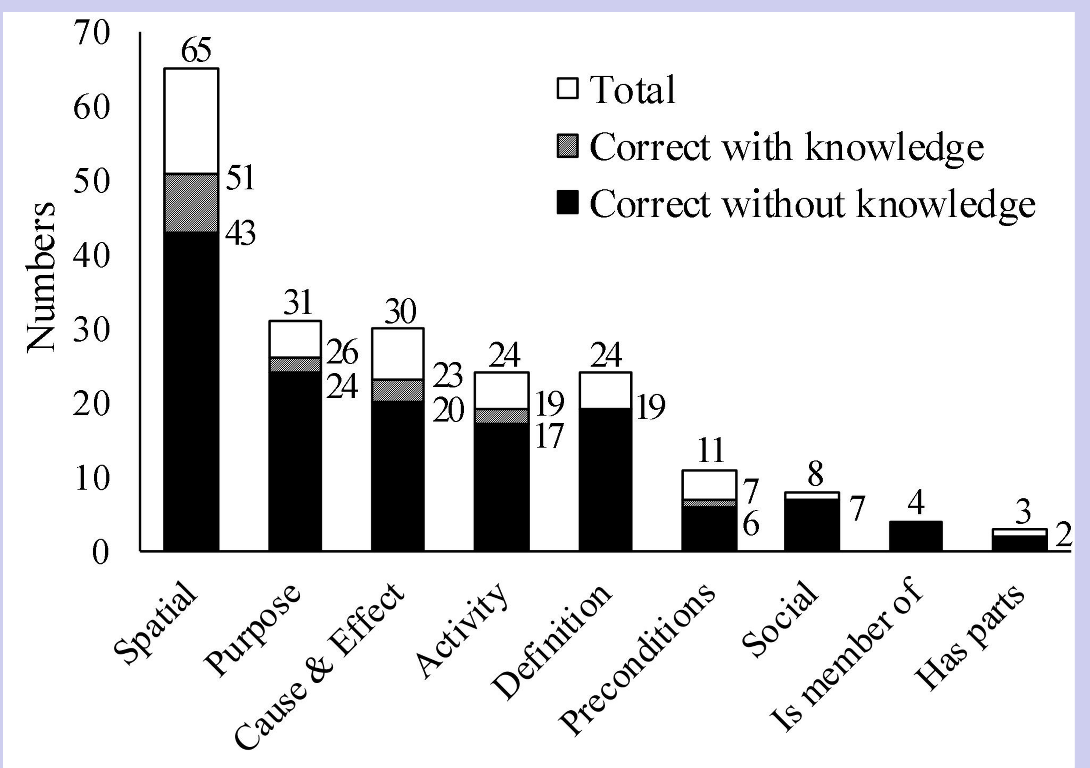
Knowledge-to-text transformation is effective and robust for knowledge-enhanced CQA.

**Answer to Q1**: By incorporating golden external knowledge, CQA can be significantly improved and can achieve close-to-human performance.

**Answer to Q2**: The potential of knowledge is still far from being fully exploited by current knowledge-enhanced CQA methods:
1) Current knowledge-enhanced CQA methods only exploit knowledge to a limited extent.
2) Despite the effectiveness of our method, there is still great potential in generating accurate question-relevant knowledge descriptions.
3) The commonsense knowledge embedded in current pretrained language models is still not enough for CQA.

| Models | WSC | HellaSWAG | SOCIAL IQa |
|---|---|---|---|
| BERT | 66.0 | 42.3 | 66.2 |
| +Knowledge | **68.1** | **44.2** | **68.8** |
| RoBERTa | 81.4 | 82.5 | 74.3 |
| +Knowledge | **82.5** | **83.0** | **75.0** |
| ALBERT | 84.9 | 86.1 | 77.2 |
| +Knowledge | **87.0** | **86.9** | **77.8** |
| Human | 92.1 | 94.5 | 86.9 |

**Knowledge is effective for CQA**:
- Knowledge can significantly improve skills including "Spatial", "Cause & Effect" ,"Activity" and "Purpose".
- For "Definition", "Social", and "Has parts" skills, the knowledge-enhanced model achieves similar performances.




## 4. Conclusions (Answer to Q3)

(1) Context-sensitive knowledge selection is critical for knowledge-enhanced CQA.
(2) The knowledge-text heterogeneity is a critical bottleneck for exploiting the information from both knowledge and text.
(3) It is valuable to incorporate more commonsense in pretrained language models.