$\texttt{E-KAR}$: A Benchmark for Rationalizing Natural Language Analogical Reasoning

Anonymous ACL submission

Abstract

The ability to recognize analogies is fundamental to human cognition. Existing benchmarks to test word analogy does not reveal the underneath process of analogical reasoning of neural models. Holding the belief that models capable of reasoning should be right for the right reasons, we propose a first-of-its-kind Explainable Knowledge-intensive Analogical Reasoning benchmark (**E-KAR**). Our benchmark consists of 1,665 problems sourced from the Civil Service Exams, which require intensive background knowledge to solve. Besides, we design a free-text explanation scheme to explain how an analogy is drawn, and manually annotate **E-KAR** with 8,325 knowledge-rich sentences of such explanations. Empirical results suggest that this benchmark is very challenging to some state-of-the-art models for both explanation generation and analogical question answering tasks, which invites further research in this area.[1]

1 Introduction

Analogy holds a vital place in human cognition, driving the discovery of new insights and the justification of everyday reasoning (Johnson-Laird, 2006; Gentner and Smith, 2012; Bartha, 2013; Bengio et al., 2021). Due to their unique value in many fields such as creativity (Goel, 1997) and education (Thagard, 1992), analogy and analogical reasoning have become a focus in AI research. The grand question is, are artificial neural networks also capable of recognizing analogies?

Relatively little attention has been paid in NLP to answer this question. The problem of recognizing analogies is mainly benchmarked in the form of (A:B::C:D) (Turney et al., 2003; Mikolov et al., 2013b; Gladkova et al., 2016; Li et al., 2018a) and targeted for testing the ability of pre-trained word embeddings. Given a tuple of terms as *query* (e.g., `tea:teapot:teacup`) and a list of *candidate answers* as in Figure 1, a model needs to find the most analogous candidate to the query, which is C in the example since it matches the relations inherent in the query better than others.

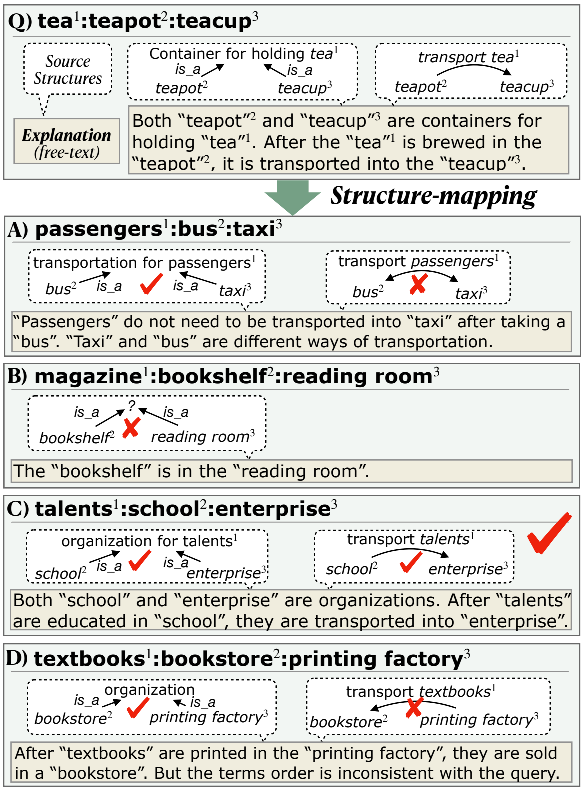


Figure 1: An example in **E-KAR**. The explanations in **E-KAR** explain the *structure-mapping* process for analogical reasoning, where source structures are drawn from the query and mapped onto each candidate answer for decision-making.

Most methods (Mikolov et al., 2013a; Levy and Goldberg, 2014; Pennington et al., 2014) hold a connectionist assumption (Feldman and Ballard, 1982) of *linear analogy* (Ethayarajh et al., 2019), that the relation between two words can be estimated by vector arithmetic of word embeddings. For example, $\vec{\texttt{king}} - \vec{\texttt{man}} + \vec{\texttt{woman}} = \vec{\texttt{queen}}$.

[1] Data will be released upon the publication of this paper.

However, current benchmarks focus on the recognition of binary analogies such as syntactic, morphological and direct semantic (e.g., *is_a* and *synonym_of*) relations. And the analogical reasoning procedure behind them is far beyond the scope of this line of research.

However, how to explain and rationalize analogical reasoning remains to be the major challenge. Psychological literature (Gick and Holyoak, 1983; Gentner, 1983; Minnameier, 2010) suggests that analogical reasoning follows the *structure-mapping* process. That is, a target (the domain where a problem must be solved, i.e., candidates) and a source (the domain where the analogy is drawn, i.e., the query) are matched, and the relevant features of the source have to be mapped onto the target. In Figure 1, source structures are drawn from the query and mapped onto candidates, where A, B, D all fail at certain structures. We argue that such a process can be verbalized into natural language to explain analogical reasoning.

Moving from simply recognizing analogies to exploring human-like reasoning for neural models, we emphasize the importance of a new kind of analogical reasoning benchmark. To fill in this blank, we propose a first-of-its-kind benchmark for **E**xplainable **K**nowledge-intensive **A**nalogical **R**easoning (**E-KAR**). We collect 1,665 analogical reasoning problems sourced from the publicly available Civil Service Examinations of China, which are challenging and knowledge-rich multiple-choice problems designed by domain experts. To justify the reasoning process, we follow the aforementioned guidelines from psychological theories and manually annotate explanations for each query and candidate answers in **E-KAR**. Since the annotation requires intensive involvement of knowledge and reasoning, we carefully design a *double-check* procedure for quality control. In summary, the contributions of this paper include:

- We advance the traditional setting of word analogy recognition by introducing a knowledge-intensive analogical reasoning benchmark (**E-KAR**), which is first-of-its-kind and challenging.
- To justify the analogical reasoning process, we design free-text explanations according to theories on human cognition, and manually annotate them.
- We define two tasks in **E-KAR**, i.e., analogical QA and explanation generation, and report the performance of some state-of-the-art neural models. We discuss the potentials of this benchmark and hope it facilitates future research on analogical reasoning.

2 Related Work

Word Analogy Recognition in NLP Benchmarks for word analogy recognition (Turney et al., 2003; Mikolov et al., 2013b; Gladkova et al., 2016; Li et al., 2018a) examine mostly linear relations between words (Ethayarajh et al., 2019). Such analogies can often be effectively solved by vector arithmetic for neural word embeddings, such as Word2Vec (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014). Recent studies (Brown et al., 2020; Ushio et al., 2021) also test such ability of pre-trained language models (PLMs) (Radford et al., 2019; Devlin et al., 2019; Brown et al., 2020) on these benchmarks. An exceptional benchmark is Li et al. (2020), where they build a knowledge-enhanced analogy benchmark that leverages word sense definitions in a commonsense knowledge base (Ma and Shih, 2018). However, these benchmarks are mainly set up for evaluating learned representations, and few of them ever investigated the analogical reasoning skills for neural models. Thus, the goal of this work largely differs from this line of research, as we aim to build a knowledge-intensive benchmark to teach neural models analogical reasoning for correct thinking.

Reasoning Benchmarks from Examinations There are abundant benchmarks derived from human examinations to facilitate the study of machine reasoning (Clark et al., 2016; Schoenick et al., 2017). For example, RACE (Lai et al., 2017) is collected from the English exams for middle and high school students, focusing on skills of passage summarization and attitude analysis. ARC (Clark et al., 2018) contains natural, grade-school science questions authored for human tests. MCQA (Guo et al., 2017), GeoSQA (Huang et al., 2019) and GCRC (Tan et al., 2021) are sourced from national college entrance exams of China, measuring a comprehensive set of reasoning abilities. LogiQA (Liu et al., 2020a) consists of logical reading comprehension problems from Civil Service Exams of China, which is also our source of analogical problems. ReClor (Yu et al., 2020) and LR-LSAT (Wang et al., 2021), collected from Law School Admission Test, aim for testing logical reasoning abilities. In our work, we focus on analogical reasoning skills for machines and additionally equip **E-KAR** with annotated explanations to rationalize reasoning.

Explainable NLP Datasets One of the most prominent objectives in machine reasoning is giving reasons or explanations for a prediction. In current datasets for explainable NLP, such reasons can be categorized into three classes (Wiegreffe and Marasović, 2021): 1) highlights explanations (Camburu et al., 2018; Yang et al., 2018; Thorne et al., 2018; Kwiatkowski et al., 2019), which are subsets of the input elements to explain a prediction, e.g., words or sentences; 2) free-text explanations (Camburu et al., 2018; Zellers et al., 2019; Aggarwal et al., 2021) that are textual explanations for justification; 3) structured explanations (Mihaylov et al., 2018; Khot et al., 2020; Clark et al., 2020; Jhamtani and Clark, 2020; Geva et al., 2021), which are not fully free-text and generally follow certain structures such as a chain of facts. The explanations can be utilized to augment (Rajani et al., 2019), supervise (Camburu et al., 2020) and evaluate (DeYoung et al., 2020) the predictions of neural models. In this work, we phrase analogical reasoning itself as an instance of machine reasoning tasks, advancing the research on analogical reasoning from the perspectives of data collection.

3 Explainable Analogical Reasoning

In this work, we consider a classic setting of analogical reasoning within NLP: recognizing word/term analogies.[2] This task can be formulated as multiple-choice question-answering. Given a query tuple Q with k (two or three) terms, and m candidate answer tuples $A = \{A_i\}_{i=1}^{m}$, the goal is to find the most analogous one in the candidates to the query.

We advocate that reasoning is about giving reasons explaining a prediction. In order to teach machines to analogize as humans do, we draw inspiration from theories in cognitive psychology to design the forms of explanations.

3.1 Analogical Reasoning: A Psychological Perspective

Before designing suitable forms of explanations, we introduce some important theories from cognitive psychology for a better understanding of analogical reasoning. In the psychological literature, analogical reasoning is described as a *schema-induction* (Gick and Holyoak, 1983) or *structure-mapping* (Gentner, 1983) process. Peirce (1896) claimed that analogy is a combination of abductive and inductive reasoning. Minnameier (2010) further developed the inferential process of analogy into three steps, which we take as the guidelines for designing explanations:

1. A possibly suitable structure in the source domain is abduced from the target domain, which might also work for the target problem;
2. The specific concepts of the source structure have to be replaced by suitable target concepts (by an inductive inference);
3. The validity of the transformation is judged w.r.t. solving the target problem.

Take Figure 1 for example. Source structures can be abduced that both term 2 (`teapot`) and term 3 (`teacup`) belong to a concept, and term 1 (`tea`) can be transported from term 2 to term 3. The mapping naturally reveals the validity, for example, candidate A is wrong because `passengers` do not follow a unidirectional transportation (i.e., from `bus` to `taxi`) but a bidirectional one.

3.2 Explanations for Analogical Reasoning

Following the above guidelines, the explanations for the analogical reasoning task should also include three parts: *1)* description of suitable structures for the query; *2)* how the structure is mapped into candidates; and *3)* reasons to justify whether the mapping is correct, such as commonsense knowledge. To this end, we define *free-text explanation* for analogical reasoning, which is one of the most expressive and commonly-used explanations (Wiegreffe and Marasović, 2021). We ensure the free-text explanations to be self-contained, knowledge-rich, and sufficient to solve the problem as a substitute for the original input.

Specifically, for each query (Q) and candidate (A_i), we define free-text explanations $\mathcal{E}_Q$ and $\mathcal{E}_{A_i}$. Following the guidelines in § 3.1, $\mathcal{E}_Q$ should describe the best suitable inherent structure in a query. $\mathcal{E}_{A_i}$ should decide the correctness of candidate A_i and provide facts as support evidence. Note that the decision should be drawn by mapping candidate terms into the structure expressed in $\mathcal{E}_Q$ correspondingly, which is analogous to template-filling.

4 The E-KAR Benchmark

Previous benchmarks consider recognizing word analogies as testbeds for evaluating pre-trained

[2] Here, "term" corresponds to "word" in previous analogy benchmarks, but allows for multiple words.

Dataset	Data Size (train / val / test)	# of Terms in Cand.	# of Cand.
SAT	0 / 37 / 337	2	5
Google	0 / 50 / 500	2	4
BATS	0 / 199 / 1,799	2	4
E-KAR	1,174 / 171 / 320	$2_{(64.7\%)}$, $3_{(35.3\%)}$	4

Table 1: Comparison between **E-KAR** and previous analogy benchmarks: data sizes in different splits, number of terms in a query or candidate answer, and number of candidates for multiple-choices.

Corpus	$n = 1$ (3.9%)	$n = 2$ (59.3%)	$n = 3$ (14.0%)	$n \geq 4$ (22.8%)	All (100%)
Ency.	88.39	95.70	85.14	73.26	88.83
Thes.	99.57	86.04	42.69	38.69	69.71
Both	100	96.15	85.73	73.33	89.64

Table 2: Proportion of terms with various number of Chinese characters (n) in the dataset, as well as their coverage (%) in different corpora (encyclopedia and thesaurus).

word embeddings. In this work, we take a step forward and build a new kind of benchmark **E-KAR** to facilitate the study of analogical reasoning.

4.1 Dataset Collection

We build our dataset upon the publicly available questions of Civil Service Exams of China (CSE), which is a comprehensive test for candidates' critical thinking and problem-solving abilities. CSE consists of problems that test various types of reasoning skills, such as graphical reasoning, logical reasoning and comprehension (Liu et al., 2020b), analogical reasoning, etc.

We collect in total 1,665 analogical reasoning problems from CSE over the years. One of the prominent features in CSE problems is the intensive involvement of commonsense, encyclopedic, and idiom knowledge. For example, one needs to be aware of the commonsense that "the `tide` is caused by both `Lunar gravity` and `Solar gravity`". More importantly, one needs to know a *negated fact* in order to reject a candidate, such as the fact that "`husband` is *not* a `job`" or "a `car` is *not* made of `tires`". We keep mainly those requiring knowledge and logical reasoning skills. The rest is manually removed, such as ones testing mathematics, morphology, and phonics, as well as the problems with terms larger than three.

Each problem consists of a query term tuple and *four* candidate answer tuples of terms, as shown in Figure 1. The dataset is randomly split into training, development, and test set at the ratio of 7:1:2. We compare **E-KAR** with previous benchmarks in Table 1, including SAT (Turney et al., 2003), Google (Mikolov et al., 2013b) and BATS (Gladkova et al., 2016). There are 35.3% problems with three terms in **E-KAR**, whereas previous ones only consist of two, making **E-KAR** more challenging.

Corpus with Background Knowledge We further build a corpus to aid the understanding of terms like idioms and rare ones that current neural networks struggle to comprehend. The corpus is built upon an encyclopedia[3] and a thesaurus[4], which are both one of the largest and most widely-used Chinese sources of their kind. Detailed statistics of coverage are reported in Table 2. Overall, the corpus covers 89.64% of all terms in **E-KAR**, showing its richness for knowledge coverage.

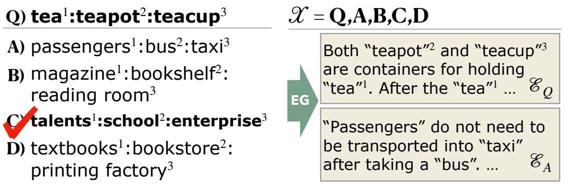
4.2 Manual Annotation of Explanations

We work with a private company for annotating the explanations defined in § 3.2. Before annotation starts, we conduct a training session for all annotators to fully understand the requirements and pick the capable ones based on a selection test. The selected workers are allocated into two teams, a team of explanation constructors and a team of checkers, where the checkers achieves better scores in the test. All of them are paid above the local minimum wage. The annotation consists of two stages: *1)* the construction stage for writing explanations, and *2)* the double-check stage for quality control.

Construction During annotation, each problem is assigned to a constructor to build five sentences of explanations: one for query and four for candidate answers. The explanations are required to be: *1)* fluent and factually correct, *2)* able to solve the problem on their own, and *3)* knowledge-rich. To reduce the labeling difficulty, we offer them sentences from the retrieved corpus for reference, while allowing them to use the search engine for querying the Internet.

First-round Checking Afterward, a problem with five annotated explanations is fed to a checker for a first-round checking. The checker decides

[3] Baidu Encyclopedia (https://www.baike.baidu.com).
[4] Xinhua Chinese Dictionary (https://www.zdic.net).



(a) Analogical QA. (b) Explanation Generation.

Figure 2: Examples of two shared tasks.

whether to accept an explanation sentence according to the criteria in the construction stage. The rejected ones are sent back to the construction team for revision along with reasons to reject, which serves to re-train the construction team. The process repeats until a batch reaches 90% accuracy. Then, a second-round checking initiates.

Second-round Checking A verified batch is presented to authors for double-checking. Authors conduct random inspections, and unqualified annotations are sent back with reasons to the check team to fine-tune their checking criteria, which in turn regularize the construction team. The process also repeats until a batch reaches 95% accuracy.

In the end, the authors manually calibrate every explanation and acquire 1,665 analogical problems and a total number of 8,325 (5×1,665) free-text explanations, with an average of 31.9 characters per sentence.

4.3 Shared Tasks in E-KAR

We define two shared tasks, *explanation generation* (EG) and *multiple-choice question-answering* (QA) for teaching models how to analogize. We denote input as $\mathcal{X} = (Q, A)$, output as $\mathcal{Y}$, and explanations as $\mathcal{E}$. Thus, the tasks can be formulated as $P_{\mathrm{EG}}(\mathcal{E}|\mathcal{X})$ and $P_{\mathrm{QA}}(\mathcal{Y}|\mathcal{X})$. Figure 2 shows the examples of input and output.

Task 1: Analogical QA As introduced in § 3, the analogical QA is be formulated as $P_{\mathrm{QA}}(\mathcal{Y}|\mathcal{X})$. The QA task requires an understanding of the relationship between the query and each of the candidates to find the correct answer. For evaluation, we directly use the *accuracy* of multiple-choice QA.

Note that all candidates may be related to the query tuple from certain perspectives, the challenge lies in finding the *most* related one. That is, we have to identify the inherent connections and relations between terms in the query and candidates, considering properties such as linguistic features, meaning, and order of terms, commonsense knowledge, etc. For example, the error for candidate D in Figure 1 can be attributed to the incorrect term order, though three terms follow a similar commonsense relationship as seen in the query. Hence, the best choice is C.

Task 2: Explanation Generation This task aims to produce the intermediate reasoning process of analogical reasoning as seen in Figure 2(b), formulated as $P_{\mathrm{EG}}(\mathcal{E}|\mathcal{X})$. Such explanations serve as training supervisions to explain and improve model predictions. As defined in § 3.2, we aim to generate $\mathcal{E}_Q$ and $\mathcal{E}_{A_i}$ for each query and candidate answer, where the former serves as the abduced source structures to be mapped onto the latter. The generated text can be evaluated with text generation metrics such as ROUGE (Lin, 2004), MoverScore (Zhao et al., 2019), BERTScore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020). However, great challenges remain for automatically evaluating semantic-rich text (Celikyilmaz et al., 2020).

5 Methods

We evaluate some of the state-of-the-art neural models on both tasks of **E-KAR**. The implementation details are reported in Appendix A.

5.1 Baselines for Analogical QA

Pre-trained Methods As pre-trained-only baselines, we adopt three static word embeddings that have shown their effectiveness in previous analogy tasks: Word2Vec (Mikolov et al., 2013a), GloVe (Pennington et al., 2014) and FastText (Bojanowski et al., 2017). We also test contextualized embeddings from PLMs, including BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). The averaged token representation is taken as the term representation. A query or a candidate is calculated as the sum of the representations of each term pair, which is represented as the embedding vector differences (Hakami and Bollegala, 2017; Ushio et al., 2021). The candidate with the highest cosine similarity to the query is chosen as the predicted answer.

Fine-tuned Methods We also set up fine-tuned baselines with PLMs (BERT and RoBERTa). Since previous benchmarks do not have a training set, we only fine-tune the models on their development set. The query and candidates are respectively *verbalized* into text using simple prompts such as

"A:B::C:D::E:F". Each candidate is concatenated with the query into one sentence, which is fed into a PLM for contextualized representation learning. Then, averaged hidden states are fed to an MLP layer and a softmax layer for classification. Besides, the semantics of terms in the problem can be enriched with background knowledge $\mathcal{K}$ from the corpus. Given a term, we retrieve the first knowledge sentence from the corpus, and concatenate it to the original input. The parameters are fine-tuned during training.

Human Evaluation We also ask three undergraduate and graduate students to solve the randomly sampled 200 problems without any hints, and report the averaged score of them as human performance.

5.2 Baselines for Explanation Generation

We formulate the EG task in a sequence-to-sequence (Seq2Seq) paradigm. Although the explanation is individually specific to each query and candidate, the generator has to take into account the whole problem for generating with the *best* source structure (as in § 3.1) and thus finding the most analogous candidate. Thus, we feed into the model the concatenation of the query and all candidates, and the model is trained to generate different explanations by changing the prefixes, e.g., "Generate: Q/A_i". The Seq2Seq model is instantiated with state-of-the-art pre-trained language models for Seq2Seq tasks, including BART (Lewis et al., 2020) and T5 (Raffel et al., 2020).

6 Results and Analysis

In this section, we wish to answer three questions: *Q1)* Can models do knowledge-intensive analogical QA? *Q2)* Can models generate rational reasons for analogical thinking? *Q3)* How do different hints help humans solve analogical problems?

Categorization of Problems We first manually categorize the relational types of problems in **E-KAR** according to a pre-defined schema. Note that, unlike free text, we are unable to induce a comprehensive set of relations that covers all candidates due to the complexity of CSE problems. As a result, we carefully assign at least one relation to each query. To facilitate analysis, we also try to assign relations to each candidate and query *in the development and test set*, ending up covering 76% of the candidates and 100% of the queries.

We refer to several sources of word analogy definitions and textbooks for analogy tests (listed in Appendix B), and categorize the relations into five *meta-relations* (as well as their coverage in the test set) and several accompanying *sub-relations*:

1. *Semantic* (R1, 8.88%), the similarity or difference in the meaning of terms, including *synonym_of* and *antonym_of*;
2. *Extension* (R2, 41.60%), the relation between the extension of terms, including *is_a*, *contradictory_to*, etc.;
3. *Intension* (R3, 34.83%), terms relate to each other by inherent properties, including *made_of*, *has_function*, etc.;
4. *Grammar* (R4, 7.74%), the grammatical relations between terms, including *subject-predicate*, *head-modifier*, etc.;
5. *Association* (R5, 6.95%), logical association between terms, including *result_of*, *sufficient_to*, etc.

Complete sub-relations are presented in Appendix B, as well as their definitions and examples.

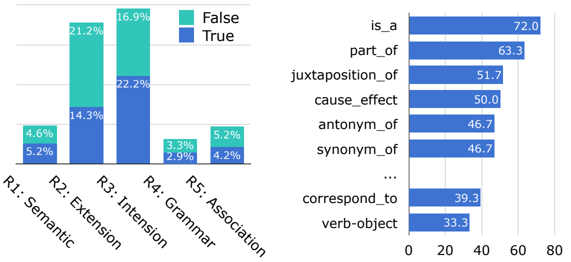
6.1 Can models do knowledge-intensive analogical QA?

Table 3 reports the accuracy results of baseline methods on previous analogy tasks and the QA task in **E-KAR**. We find that contextualized word embeddings from PLMs are not very competitive against static word embeddings in previous analogy tasks, which is consistent with the findings in Peters et al. (2018). In more knowledge-rich datasets such as **E-KAR**, the opposite conclusion can be made, with PLMs prevailing over static word embeddings. Also, humans achieve 77.8% accuracy in **E-KAR**, indicating the challenge of this task as well as showing that neural models still fall far behind human performance.

Performance from contextualized representations can be improved in all tasks through fine-tuning, especially for **E-KAR**, where accuracy increases by roughly 5 to 6 points. When augmented with knowledge from corpus through naïve sentence concatenation, however, the accuracy drops considerably. This is probably because the first sentence of a term in the corpus only describes limited properties of the term itself, but analogical reasoning requires the deep understanding of the relationship between the terms. Also, with the concatenation of knowledge sentences, longer input distracts a model from solving the problem. We

Method	SAT	Google	BATS	E-KAR
Pre-trained Word Embeddings				
Word2Vec†	41.5	93.2	63.9	28.2
GloVe†	**47.7**	96.0	67.6	30.9
FastText†	47.1	**96.6**	**72.0**	**31.4**
Pre-trained Language Models				
$\text{BERT}_b^\dagger$	32.9	80.8	61.5	34.5
$\text{RoBERTa}_b^\dagger$	42.4	90.8	69.7	41.7
$\text{RoBERTa}_l^\dagger$	**45.4**	**93.4**	**72.2**	**44.6**
Fine-tuned Language Models				
BERT_b	38.9	86.6	68.0	41.8
RoBERTa_b	47.7	93.8	75.2	46.9
RoBERTa_l	**51.6**	**96.9**	**78.2**	**50.1**
$+\mathcal{K}$	-	-	-	44.2
$+\mathcal{E}$	-	-	-	95.0
Humans	-	-	-	77.8

Table 3: Accuracy results on previous analogy tasks and the QA task in **E-KAR**. Method† is not tuned. PLM_b or PLM_l denote *base* or *large* version respectively. Method + $\mathcal{K}$ and $\mathcal{E}$ denote the input is concatenated with retrieved knowledge and gold explanations respectively.



(a) Meta-relations distributions and their error ratios. (b) Sub-relations in a sorted order of error rate.

Figure 3: Error analysis of different query relations. The results are predicted by a fine-tuned RoBERTa (large).

Method	RG2.	Mover.	BERT.	BLRT.
T5_b	30.80	**64.55**	**76.33**	60.34
BART_b	33.04	64.30	70.78	62.16
BART_l	**34.49**	64.40	71.26	**63.15**

Table 4: Results of different explanation generation models w.r.t. ROUGE-2, MoverScore, BERTScore and BLEURT.

believe a more delicate way of knowledge injection in this task is worth investigating in the future. Notably, gold explanations help boost the accuracy of a RoBERTa model from 50.1% to 95.0%, showing good quality.

Error Analysis We further conduct an error analysis based on the results in **E-KAR** predicted by fine-tuned RoBERTa (large). The erroneous ones are classified based on the manually annotated meta-relations and sub-relations of *queries*, which are fine-grained analysis tools for a model's predictions. Figure 3(a) shows that the model perform evenly bad on all meta-relations, with R2 (Extension) being the most error-prone one (only 40.3% accuracy) and R3 (Intension) being the least one (56.8% accuracy). Figure 3(b) presents the error rate of finer-grained sub-relations with more than 10 cases. We find that, consistent with Figure 3(a), the three most error-prone sub-relations *is_a*, *part_of* and *juxtaposition_of* all belong to R2 (Extension). Besides, the model seems to do well in linguistic knowledge, with *verb-object* achieving only 33.3% error rate. These findings may shed light on future directions for knowledge-injection and reasoning with language models.

6.2 Can models generate rational reasons for analogical thinking?

We report the automatic evaluation results of generated explanations in Table 4. However, such results hardly mean anything due to the incapability to evaluate semantic-rich text of current automatic metrics. Therefore, we also randomly select 100 sentences generated by a BART (large) for manual inspection. Interestingly, we find the generated explanations do not contain much of the negated facts, which are important to refute a candidate, as mentioned in § 4.1. For explanations of refuted candidates, we find ∼90% gold ones contain negated facts for deciding correctness. However, the number drops to ∼23% in the generated ones. An interesting conclusion can be drawn that current generative models do not seem to know how to generate a negated fact which is still truthful, such as "`feeling` can *not* guide `psychological reaction`." since feeling *is* a reaction.

The fact also questions the astonishing performance boost (from 50.1% to 95.0%) in QA by gold explanations, as it could be biased towards surface-level negation. To debias this, we conduct a simple ablation study by directly removing the clauses containing the negation word "不"(*not*) in the test set, and still achieve 90.9% in QA accuracy. These findings point to the potential of a high quality analogical reasoning system given correct generated explanations.

To sum up, the errors for generated explanations can be roughly categorized into three classes: 1) incapable of generating negated facts; 2) generating

Q)	氧气 (oxygen):臭氧 (ozone)
A)	盐 (salt):氯化钠 (sodium chloride)
B)	硫酸 (sulfuric acid):硫 (sulfur)
C)	石墨 (graphite):金刚石 (diamond)
D)	石灰水 (lime water):氢氧化钙 (calcium hydroxide)
$\mathcal{E}_Q^{\dagger}$	氧气和臭氧都只由氧元素组成。Both oxygen and ozone are made of only the oxygen element.
$\mathcal{E}_Q^{\ddagger}$	臭氧是氧气的一种。Ozone is a kind of oxygen.
$\mathcal{E}_A^{\dagger}$	氯化钠是盐的主要成分，盐和氯化钠不是只由一种元素组成。Sodium chloride is the main component of salt. Neither salt nor sodium chloride is made of only one element.
$\mathcal{E}_A^{\ddagger}$	氯化钠是盐的一种。Sodium chloride is a kind of salt.

Table 5: Case study of explanations, where $\mathcal{E}^{\dagger}$ is gold explanation and $\mathcal{E}^{\ddagger}$ is generated by a BART (large).

Hint	Helpfulness		
	Not (0)	Some (1)	Very (2)
Retrieved $\mathcal{K}$	45.4%	45.3%	9.3%
Expl. (Generated)	34.6%	50.7%	14.7%
Expl. (Gold)	0.0%	5.3%	94.7%

Table 6: Human evaluation on the helpfulness of different hints for solving problems in **E-KAR**.

factually incorrect statements; 3) biasing towards common patterns, such as "`term 1` and `term 2` have similar meanings" and "`term 1` is a `term 2`". For example, in Table 5, both generated $\mathcal{E}_Q$ and $\mathcal{E}_A$ are factually incorrect, and BART fails to generate the negated fact that "both are not exclusively made of one component."

6.3 How do different hints help humans solve analogical problems?

We acknowledge the limitation of automatic evaluation for explanation generation and knowledge retrieval. Therefore, we hope to figure out how background knowledge and different explanations help humans solve analogical problems.

We ask three graduate and undergraduate students as participants to complete randomly sampled 150 analogical problems. The participants are exposed with *three* settings of hints (i.e., 50 problems per setting): 1) retrieved knowledge, 2) generated explanations by a BART (large), and 3) gold explanations. Participants are asked to rate each hint based on the degree of difficulty it reduces when thinking, including unhelpful (0), somewhat helpful (1, answers can be drawn partly from hints), and very helpful (2, answers can be largely drawn from hints).[5]

According to Table 6, the gold explanations undoubtedly is the most helpful hint among them, showing its good quality. The generated explanations receives 50.7% votes of somewhat helpful (1) and 14.7% votes of very helpful (2). The retrieved knowledge achieves the worst performance in helpfulness, which can be attributed to the fact that the retrieval is purely off-the-shelf. Still, more than a half cases of retrieved knowledge (54.6%) are decided to be helpful to different extent.

7 Conclusion

In this work, we propose a first-of-its-kind benchmark **E-KAR** for explainable analogical reasoning, which sets a concrete playground and evaluation benchmark to boost the development of human-like analogical reasoning algorithms. The **E-KAR** benchmark is featured by its rich coverage in knowledge and well-designed free-text explanations to rationalize analogical reasoning process.

However, there are still many open questions that need to be addressed. For example, humans solve the analogical problems in a trial-and-error manner, but the annotated explanations in **E-KAR** are mostly post-hoc and reflect only the final step of the reasoning. Such explanations cannot offer supervision for intermediate reasoning, though it is an interesting question whether an intelligent model should be deeply supervised at every step (Tafjord et al., 2021). Furthermore, **E-KAR** only presents one feasible explanation for each problem, whereas there may be several.

This benchmark also invites analogical reasoning models that can effectively interact with extra knowledge as well as better metrics for evaluating free-text explanations. It remains to be a great challenge to generate factually correct explanations as well as negated facts. Especially, the latter is relatively under-explored in the research community but of much importance. Finally, whether the analogical QA system can correctly exploit explanations and background knowledge is also worth investigating, which may intersect with researches on debiasing (Tang et al., 2020; Niu et al., 2021).

We hope this dataset to be a valuable supplement to future research on natural language reasoning, especially for researches on analogical reasoning and explainable NLP.

[5] They reach moderate inter-rater agreement with Fleiss' $\kappa = 0.427$.

Ethical Considerations

This paper proposes a new kind of analogical benchmark with explanations to rationalize models' predictions. The dataset is collected from Civil Service Exams of China, which is publicly available and has been used in other public datasets before, such as LogiQA (Liu et al., 2020a). The annotated explanations for each problem in our dataset are crowd-sourced by working with a private company. The construction team remains anonymous to the authors, and the annotation quality is guaranteed by the double-check strategy as mentioned in § 4.2. We ensure that all annotators' privacy rights are respected in the annotation process. All annotators have been paid above local minimum wage and consented to use the datasets for research purposes covered in our paper.

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New Dataset and Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, Online. Association for Computational Linguistics.

Paul Bartha. 2013. Analogy and analogical reasoning.

Yoshua Bengio, Yann Lecun, and Geoffrey Hinton. 2021. Deep learning for ai. *Commun. ACM*, 64(7):58–65.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. Make up your mind! adversarial generation of inconsistent natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4157–4165, Online. Association for Computational Linguistics.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as Soft Reasoners over Language. pages 3882–3890.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Towards understanding linear word analogies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3253–3262, Florence, Italy. Association for Computational Linguistics.

Jerome A Feldman and Dana H Ballard. 1982. Connectionist models and their properties. *Cognitive science*, 6(3):205–254.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170.

Dedre Gentner and Linsey Smith. 2012. Analogical reasoning. *Encyclopedia of human behavior*, 2:130–136.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Mary L Gick and Keith J Holyoak. 1983. Schema induction and analogical transfer. *Cognitive psychology*, 15(1):1–38.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California. Association for Computational Linguistics.

Ashok K Goel. 1997. Design, analogy, and creativity. *IEEE expert*, 12(3):62–70.

Shangmin Guo, Xiangrong Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2017. Which is the effective way for gaokao: Information retrieval or neural networks? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 111–120.

Huda Hakami and Danushka Bollegala. 2017. Compositional approaches for representing relations between words: A comparative study. *Knowledge-Based Systems*, 136:172–182.

Zixian Huang, Yulin Shen, Xiao Li, Gong Cheng, Lin Zhou, Xinyu Dai, Yuzhong Qu, et al. 2019. Geosqa: A benchmark for scenario-based question answering in the geography domain at high school level. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5866–5871.

Harsh Jhamtani and Peter Clark. 2020. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 137–150, Online. Association for Computational Linguistics.

Philip Nicholas Johnson-Laird. 2006. *How we reason.* Oxford University Press, USA.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8082–8090.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Peng-Hsuan Li, Tsan-Yu Yang, and Wei-Yun Ma. 2020. CA-EHN: Commonsense analogy from E-HowNet. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2984–2990, Marseille, France. European Language Resources Association.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018a. Analogical reasoning on Chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143, Melbourne, Australia. Association for Computational Linguistics.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018b. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020a. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In *IJCAI*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020b. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3622–3628. International Joint Conferences on Artificial Intelligence Organization. Main track.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Wei-Yun Ma and Yueh-Yin Shih. 2018. Extended HowNet 2.0 – an entity-relation common-sense representation model. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Gerhard Minnameier. 2010. Abduction, induction, and analogy. In *Model-based reasoning in science and technology*, pages 107–119. Springer.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710.

Charles S Peirce. 1896. Lessons from the history of science. *C. Hartshorne*, 660.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Carissa Schoenick, Peter Clark, Oyvind Tafjord, Peter Turney, and Oren Etzioni. 2017. Moving beyond the turing test with the allen ai science challenge. *Communications of the ACM*, 60(9):60–64.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online. Association for Computational Linguistics.

Hongye Tan, Xiaoyue Wang, Yu Ji, Ru Li, Xiaoli Li, Zhiwei Hu, Yunxiao Zhao, and Xiaoqi Han. 2021. GCRC: A new challenging MRC dataset from Gaokao Chinese for explainable evaluation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1319–1330, Online. Association for Computational Linguistics.

Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. 2020. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *Advances in Neural Information Processing*

Systems, volume 33, pages 1513–1524. Curran Associates, Inc.

Paul Thagard. 1992. Analogy, explanation, and education. *Journal of Research in science Teaching*, 29(6):537–544.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Peter D Turney, Michael L Littman, Jeffrey Bigham, and Victor Shnayder. 2003. Combining independent modules in lexical multiple-choice problems. *Recent Advances in Natural Language Processing III: Selected Papers from RANLP*, 2003:101–110.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3609–3624, Online. Association for Computational Linguistics.

Siyuan Wang, Zhongkun Liu, Wanjun Zhong, Ming Zhou, Zhongyu Wei, Zhumin Chen, and Nan Duan. 2021. From lsat: The progress and challenges of complex reasoning. *arXiv preprint arXiv:2108.00648*.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable nlp. In *Proceedings of NeurIPS*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations (ICLR)*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zhuosheng Zhang, Hanqing Zhang, Keming Chen, Yuhang Guo, Jingyun Hua, Yulong Wang, and Ming Zhou. 2021. Mengzi: Towards lightweight yet ingenious pre-trained models for chinese.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

A Implementation Details

The pre-trained word embeddings are provided by Li et al. (2018b), and the checkpoints for PLMs by HuggingFace (Wolf et al., 2020). Most of the parameters in the baseline models take the default values from HuggingFace's Transformers library, and we keep the best checkpoint on the validation set for testing. The Chinese version of BERT (whole word masking) and RoBERTa (whole word masking extended) are provided by Cui et al. (2020), BART by Shao et al. (2021) and T5 by Zhang et al. (2021).

B Detailed Relation Definitions

For designing the relation taxonomy, we refer to a number of sources for categorizing types of analogy tests, including MAT[6], Fibonicci[7], Offcn Education (in Chinese)[8] and Huatu Education (in Chinese)[9], etc.

The complete set of meta-relations and sub-relations are presented in Table 7.

[6]http://www.west.net/s̃tewart/mat/analogies_types.htm

[7]https://www.fibonicci.com/verbal-reasoning/analogies-examples/

[8]https://www.offcn.com

[9]https://www.huatu.com

Relation	Definition	Example	Coverage
R1: ***Semantic***			**8.88%**
1) *synonym_of*	The meanings of two terms are similar.	clarity : transparency	4.48%
2) *antonym_of*	The meaning of two terms are opposite or used to express different concepts.	harmony : conflict	4.40%
R2: ***Extension***			**41.60%**
1) *identical_to*	The meanings of two terms are identical.	highway : road	3.34%
2) *is_a*	One term is the hypernym of the other.	Earth : planet	9.15%
3) *part_of*	One term is a part of the other.	steering wheel : sedan	9.32%
4) *juxtaposition_to*	Two terms belong to the same hypernym or have the same properties or functions.	shoes : socks	14.42%
5) *contradictory_to*	Two term are contradictory to each other.	vowel : consonant	0.79%
6) *contrary_to*	Two propositions cannot both be true, but can both be false.	black : white	2.55%
7) *intersection_to*	The extension of the two terms intersects.	solo : pianolude	1.67%
8) *utterly_different*	The extensions of terms do not overlap.	apple : nuts	0.35%
R3: ***Intension***			**34.83%**
1) *attribute_of*	One term is the attribute of the other.	object : inertia	1.50%
2) *probabilistic_attribute*	One term is probably the attribute of the other.	shoes : high heels	0.09%
3) *has_function*	One term has the function of the other.	calculator : calculate	4.57%
4) *metaphor*	A term is the metaphor of the other, reflecting something abstract indirectly.	pigeon : peace	1.06%
5) *takes_place_in*	A term takes place in the other.	soldier : battlefield	1.41%
6) *located_in*	A term is located in the other.	Rhine : Europe	1.50%
7) *made_of*	One term is the raw material of the other.	door : wood	3.69%
8) *tool_of*	One term is the tool of the other.	knives : murder	0.35%
9) *target_of*	One term is the target of the other.	health : exercise	0.53%
10) *corresponds_to*	Terms generally correspond to each other.	post office : mail bank	20.14%
R4: ***Grammar***			**7.74%**
1) *subject-predicate*	The originator of the action and the action itself.	plane : take off	1.32%
2) *verb-object*	The action and the object on which the action acts.	transfer : goods	3.87%
3) *head-modifier*	The preceding term modifies the other.	affluence : living	0.97%
4) *subject-object*	The originator and receiver of an action.	dairy farmer : milk	1.58%
R5: ***Association***			**6.95%**
1) *result_of*	One term causes the other.	lack of water : plants wither	3.87%
2) *follow*	The terms have a chronological or other sequential relationship, but one term does not cause the other.	sign up : take the exam	1.79%
3) *sufficient_to*	One term is a sufficient condition for the other.	raining : wet ground	0.0%
4) *necessary_to*	One term is a necessary condition for the other.	admission : graduation	1.32%

Table 7: Complete set of defined sub-relations with definitions, examples and coverage in the test set of **E-KAR**.