


Article

Semantic Representation Using Sub-Symbolic Knowledge in Commonsense Reasoning [†]

Dongsuk Oh ^{1,†} , Jungwoo Lim ^{1,‡}, Kinam Park ² and Heuseok Lim ^{1,*}¹ Department of Computer Science and Engineering, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Korea² Human-Inspired AI and Computing Research Center, Korea University, Seoul 02841, Korea

* Correspondence: limhseok@korea.ac.kr

[†] This study is an extension of our research Presented at the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020. We have extended our previous study by (1) showing how to assess pre-trained models on their understanding of questions and demonstrating language model limitations, (2) proposing a new graph representation strategy with expansion using the AMR graph and ConceptNet, and (3) showing significant performance improvements in the diverse commonsense reasoning-based datasets compared with baselines.[‡] These authors contributed equally to this work.

Abstract: The commonsense question and answering (CSQA) system predicts the right answer based on a comprehensive understanding of the question. Previous research has developed models that use QA pairs, the corresponding evidence, or the knowledge graph as an input. Each method executes QA tasks with representations of pre-trained language models. However, the ability of the pre-trained language model to comprehend completely remains debatable. In this study, adversarial attack experiments were conducted on question-understanding. We examined the restrictions on the question-reasoning process of the pre-trained language model, and then demonstrated the need for models to use the logical structure of abstract meaning representations (AMRs). Additionally, the experimental results demonstrated that the method performed best when the AMR graph was extended with ConceptNet. With this extension, our proposed method outperformed the baseline in diverse commonsense-reasoning QA tasks.

Keywords: abstract meaning representation; semantic representation; sub-symbolic; commonsense reasoning; ConceptNet; commonsense question and answering; pre-trained language model



Citation: Oh, D.; Lim, J.; Park, K.; Lim, H. Semantic Representation Using Sub-Symbolic Knowledge in Commonsense Reasoning. *Appl. Sci.* **2022**, *12*, 9202. <https://doi.org/10.3390/app12189202>

Academic Editor: Valentino Santucci

Received: 12 August 2022

Accepted: 10 September 2022

Published: 14 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Based on a clear understanding of the question and commonsense data, the commonsense question and answering (CSQA) system evaluates a question to obtain the correct answer. To predict the correct answer, a query has to be comprehensively understood with commonsense knowledge. As shown in Figure 1, *ferret* is a key word for answering the question. Unlike machines, people capture the relationships between the predicates and arguments of a question and extract the necessary concepts from commonsense knowledge. However, the machine implicitly gathers statistics on how words appear when combined with large corpora rather than obtaining a clear representation of concepts [1]. An ultimate machine capable of reasoning commonsense must understand linguistic symbols, such as semantic representations [2–5]. Furthermore, selecting the concept of the question analytically from considerable commonsense knowledge is necessary for precise reasoning.

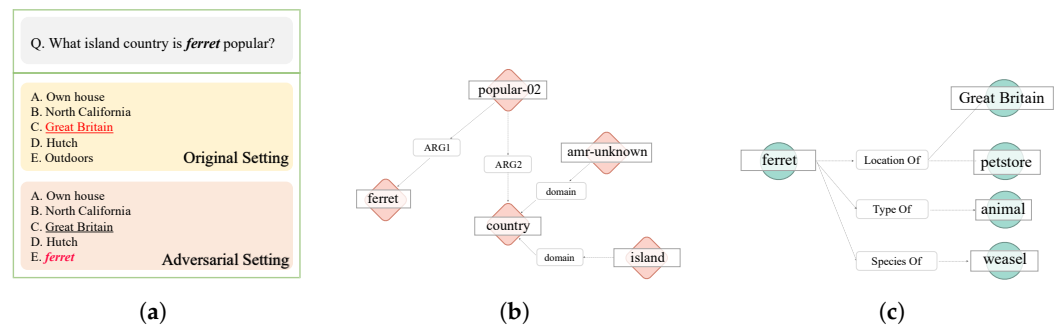


Figure 1. ((a) Adversarial Attack in Commonsense Question Answering Task) Comparison of the prediction of the BERT-base-based model between the original and adversarial settings. Bold texts indicate the randomly selected argument among the arguments of the sentences. The underlined option is the correct answer and the red colored option indicates the predicted answer from the pre-trained language model, ((b) Example of AMR graph of question) Example of AMR graph of a question, and ((c) Example of ConceptNet subgraph of question) Example of ConceptNet graph.

Recent tasks have mainly focused on answering questions given relevant documentation or context, requiring little general background knowledge. However, people use their wealth of world knowledge to answer questions. Due to the increasing demand for evaluating the capability of machines on commonsense reasoning [6,7] similar to humans, corresponding datasets appeared recently. OpenBookQA [7] is a new kind of question-and-answer dataset modeled on open book exams for assessing human comprehension of a topic. Answering the questions in this dataset requires additional extensive general knowledge not covered in the book. In addition, the CommonsenseQA dataset [6] builds using ConceptNet, a knowledge base that contains people's common sense. This dataset leverages multiple target concepts with the same semantic relationship to a single source concept in the ConceptNet. Thus, the model must distinguish each target concept from the mentioned source.

Two main approaches have been developed to solve these tasks from the model. The first approach to commonsense reasoning is a fine-tuning method with a pre-trained language model. This method collects evidence sentences from knowledge sources, such as Wikipedia or open mind commonsense (OMCS) [8], and trains a pre-trained language model using this external commonsense knowledge. During the inference stage, the system creates an input as a concatenation of questions, candidate answers, and corresponding evidence retrieved from the evidence sources. The system carelessly trains evidence sentences with commonsense knowledge using a model with numerous parameters. The second approach applies a reasoning process with a commonsense knowledge graph [9–11]. Based on the words that appear in the question, this method extracts data from ConceptNet [12] and represents it using graph encoders. The answer is predicted using graph representation and attention data from the language models. The system supplements the insufficient representations of the language model using a commonsense knowledge graph. However, improving the performance of these approaches without understanding the question remains difficult.

To improve performance, this study proposes the abstract meaning representation (AMR) of a question [13]. Within a logical framework, AMR graphs the meaning of single or multiple sentences. Because its representation lacks a commonsense relationship between concepts, we expanded this graph with an additional commonsense knowledge graph based on each concept for commonsense reasoning. We used the AMR symbolic framework to logically comprehend commonsense reasoning and represent the new AMR-ConceptNet graph, which is expanded with commonsense knowledge in a Levi graph [14].

Our main contributions are as follows.

- We demonstrate how to assess pre-trained models on the understanding of the questions and demonstrate the limitations of the language models.

- We propose a new graph representation strategy expanded with an AMR graph and ConceptNet.
- Compared with the baselines, our method shows significant performance improvement in diverse commonsense reasoning-based datasets.

2. Related Work

2.1. Abstract Meaning Representation (AMR)

Various studies in the natural language processing (NLP) fields have used AMR in their models [5,15–22]. An AMR [13] is a graph that logically represents the meaning of a sentence. The AMR graph captures the structure of “who is doing what to whom” in a given sentence and represents the sentence with a directed acyclic graph including single-rooted nodes, concept nodes, and relationships between them. Because the root node is a representation focus, the other concepts are connected to semantic relations once the root node is fixed. Similar to a parse tree, the AMR graph is traversable, considering all words. However, AMR builds the same graph structure for two syntactically different but semantically similar sentences. The concepts in an AMR graph are referred to as the events or entities. Additionally, the relationships between the concepts are denoted in the vocabulary of the PropBank frameset [23] and standard words. AMR represents semantic roles using more than 100 semantic relations (for example, negation, conjunction, and command). In PropBank, the graph form is labeled as the semantic roles of ARG0~4 and ARGM. Subsequently, other concept nodes are sequentially joined from the semantic relations.

2.2. ConceptNet

ConceptNet [12] is a multilingual knowledge graph that connects words and phrases of natural language used by people in the real world. Real-world common sense is defined in ConceptNet as two nodes and directed edges indicating concepts and their relations, respectively. The relationships defined in a single lexical resource are not enough for a machine to understand words in the natural language people use. For example, in WordNet, dog and cat are defined as hyponyms for animal. However, it is not connected to a pet. ConceptNet is constructed by collecting data from various knowledge bases, including Wiktionary [24], WordNet [25], and DBpedia [26]. They were also defined as hierarchical URLs to avoid ambiguity. For example, the node “/c/en/read/v” can be retrieved using the part-of-speech data. Moreover, multiple relationships simultaneously exist between two different nodes. Consequently, the ambiguity between two different nodes can be handled using these multiple relations.

2.3. Commonsense Reasoning

Pre-trained-based models have performed admirably in earlier studies. The initial approach (<https://gist.github.com/commonsensepretraining/507aefddcd00f891c83ebf6936df15e8> (accessed date in 1 May 2022), <https://drive.google.com/file/d/1sGJBV38aG706EAR75F7LYwCqci9ocG9i/view> (accessed date in 1 May 2022)) for commonsense reasoning was based on a fine-tuning method. This approach uses questions and answers to limit reasoning capability. A retrieval module was also used to supplement the reasoning ability to retrieve evidence from the questions and answers. The second approach uses an additional encoder to embed knowledge graphs such as ConceptNet. An additional encoder typically uses paths or nodes in the graph [9–11]. Lin et al. [9], Ma et al. [11] extracts the graph paths using a specific search algorithm and uses them as the input for the encoder. Lv et al. [10] embeds the node with an adjacency matrix and uses the graph attention to compute the attention score. Various pre-trained models that used the aforementioned methods achieved high performance, including BERT [27] and RoBERTa [28], which use bidirectional transformer encoders. They also include XLNet [29], which uses autoregressive language modeling; ALBERT [30], which uses cross-layer parameter sharing and factorized embedding parameterization; and ELECTRA [31]. Additionally, ELECTRA

was designed as a generator and discriminator, pre-trained using a replaced token detection (RTD) task.

3. Proposed Method

Because the word in a sentence acts in a specific role, such as a predicate or argument, the concept of the AMR graph also has semantics in the graph structure. Owing to these graph structure benefits, we used AMR graphs to extract commonsense knowledge graphs. We generated the question as an AMR graph using the model of Cai and Lam [32], which has recently demonstrated excellent performance in AMR generation tasks. Although most AMR graphs are properly generated from the model, inevitable errors in the types of relationships or concepts may occur. After obtaining an AMR graph, our method integrates it with the ConceptNet graph. Particularly, if the AMR concept existed in ConceptNet, it connected the ConceptNet node with all nodes in the AMR. The proposed methods then used ConceptNet to prune ARG0, ARG1, ARG2, ARG3, and ARG4 nodes that lacked edges. Our proposed graph representation prunes other nodes unrelated to the argument nodes because these argument relationships have significant meanings. This process uses ACP-ARG graphs to train the model by repeatedly identifying excessive paths. The proposed ACP-ARG is shown in Figure 2.

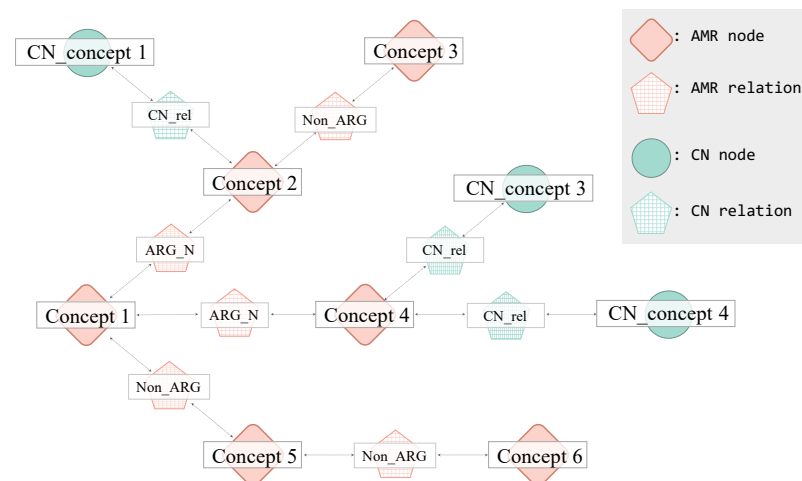


Figure 2. Example of an ACP-ARG Graph.

The graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ represents a fixed set of nodes \mathcal{V} and relational edges \mathcal{E} . The ACP-ARG graph can be expressed as follows:

$$\mathcal{G}_{ACP-ARG} = (\{\mathcal{V}_{amr} \cup \mathcal{V}_{cn}^{amr^{arg}}\}, \{\mathcal{E}_{amr} \cup \mathcal{E}_{cn}^{amr^{arg}}\}) \quad (1)$$

The union of the AMR graph and ConceptNet subgraphs containing AMR concepts linked with argument relations results in the ACP-ARG graph denoted by Equation (1). The AMR graph is denoted as $\mathcal{G}_{AMR} = \{\mathcal{V}_{amr}, \mathcal{E}_{amr}\}$. ConceptNet's subgraph corresponding to concepts linked to argument relations is denoted as $\mathcal{G}_{CN}^{AMR^{arg}} = \{\mathcal{V}_{cn}^{amr^{arg}}, \mathcal{E}_{cn}^{amr^{arg}}\}$.

4. Experiments

4.1. Data Setup

SRLAttack dataset is an adversarial attack dataset to analyze whether the pre-trained language model relies on superficial cues. Semantic role labeling (SRL) captures the relationship between the predicate and the arguments in the question. By determining the precise meaning of the question, the analysis results of the SRL can be used to determine the correct answer. First, we labeled the semantic roles of each argument using a pre-trained model (https://docs.allennlp.org/models/main/models/structured_prediction/predictors/srl/ (accessed date in 1 April 2021)) [33]. Then, we randomly selected the

predicate in the labeled question and determined the answer of the distractor from an argument based on the position of the predicate.

CommonsenseQA dataset consists of 12,102 questions and five candidate answers provided by Talmor et al. [6]. We divided the official training set for the experimental efficiency, and the organizer can only validate the official test set once every two weeks. The new training, development, and test sets contained 8500, 1221, and 1242, respectively.

OpenBookQA OpenBookQA dataset consists of one question, and four candidate answers about elementary-level science. The questions require common sense knowledge to solve and include 4957 training sets, 500 development sets, and 500 test sets [7]. Unlike CSQA, we used an official development set because the test set is accessible to the public; therefore, we could monitor the performance as needed.

4.2. Experimental Details

We trained the model using Quadro RTX 8000 and used the same parameters as in Cai and Lam [15] and Lim et al. [34]. The hyperparameters of each language model were obtained manually.

4.3. Baselines

4.3.1. Pre-Trained Language Models

BERT [27], a bi-directional model using the transformer architecture, performs admirably in most natural language understanding tasks. BERT is pre-trained using large-scale text data on masked language modeling (MLM) and next sentence prediction (NSP). It effectively captures the natural language context. Despite its capability to learn context, BERT cannot capture the overall meaning of a text using a static-masking rule based on 15% of the sentence. ELECTRA [31] proposed more efficient pre-training strategies using a generator structure and discriminator networks, similar to a generative adversarial network (GAN).

4.3.2. AMR-CN Reasoning Model

For the AMR-CN reasoning baseline, we used Lim et al. [34]’s model, which considers the pruned graph as an input and calculates the attention score of each path using the graph transformer and obtaining the entire graph vector. The model is shown in Figure 3.

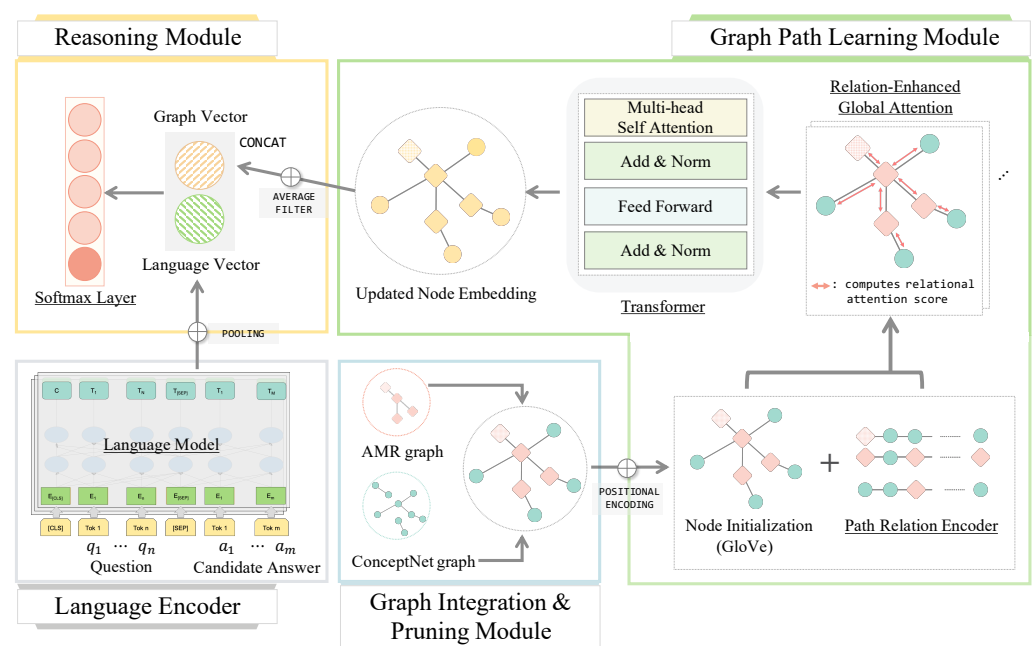


Figure 3. Overview of AMR-CN reasoning model.

4.3.3. Graph Path Learning Module

With the ACP-ARG graph from the graph integration and pruning module, the graph path learning module initializes the concept node vector as the sum of the concept embedding using GloVe [35] and absolute position embedding. A relation encoder is first used to encode the connection between the two concepts into a distributed representation for the model to recognize the explicit path of the ACP-ARG graph. The relation encoder recognizes the shortest path between two concepts and expresses the sequence as a relation vector by using a Gated Recurrent Unit (GRU) [36]. The equation representing the relation is depicted as follows:

$$\vec{p}_t = \text{GRU}_f(\vec{p}_{t-1}, sp_t), \overleftarrow{p}_t = \text{GRU}_g(\overleftarrow{p}_{t+1}, sp_t) \quad (2)$$

where sp_t is the shortest path of the relation between two nodes. The final relation encoding r_{ij} between concepts i and j is the concatenation of the final hidden states from the forward and backward GRU networks, which are represented in the Equation (3).

$$r_{ij} = [\vec{p}_n; \overleftarrow{p}_0] \quad (3)$$

In order to inject this relation information into the conceptual representation, the AMR-CN reasoning model follows the idea of relative position, including [37,38], which introduces an attention scoring method based on the conceptual representation and the relation representation.

To calculate the attention score, the model is divided the relation vector r_{ij} passed from the linear layer into forward relation encoding $r_{i \rightarrow j}$ and backward relation encoding $r_{j \rightarrow i}$, as follows:

$$[r_{i \rightarrow j}; r_{j \rightarrow i}] = W_r r_{ij} \quad (4)$$

where W_r is the parameter matrix. This split renders the model to consider the bi-directionality of the path. Thereafter, the model computes the attention score considering the concepts and their relations. Note that c_i and c_j are the concept embedding. The equation is presented below:

$$\begin{aligned} s_{ij} &= f(c_i, c_j, r_{ij}) \\ &= (c_i + r_{i \rightarrow j}) W_q^\top W_k (c_j + r_{j \rightarrow i}) \\ &= c_i W_q^\top W_k c_j + c_i W_q^\top W_k r_{j \rightarrow i} \\ &\quad + r_{i \rightarrow j} W_q^\top W_k c_j + r_{i \rightarrow j} W_q^\top W_k r_{j \rightarrow i} \end{aligned} \quad (5)$$

The first term in the last line of Equation (5) is the original term in the vanilla attention mechanism, which includes the pure content of the concept. The second and third terms capture the bias of the relation concerning the source and target, respectively. The last item represents universal relation bias. As a result, the computed attention score updates the concept embedding while maintaining a fully connected communication [15]. Therefore, the concept-relation interaction can be injected into the concept node vector. The resulting conceptual representations are aggregated into the entire graph vector and fed into the transformer layer to model the interaction between the AMR and ConceptNet conceptual representations.

The major advantage of this relation-enhanced attention mechanism is that it provides a fully connected view of input graphs using the relation multi-head attention mechanisms. By integrating two different concept types from the AMR graph and ConceptNet into a single graph, the model globally recognizes which path has high relevance to the question during the interpretation.

4.3.4. Language Encoder

The language encoder is utilized to encode text input into distributed representation, which is a pre-trained language model with a large corpus. The language model uses the models described in the baseline.

4.3.5. Reasoning Module

The proposed method performs commonsense reasoning on the ACP-ARG graph and predicts the correct answer. The model takes two types of input, text and graph representations, and transforms semantic representations into distributed representations. After obtaining text representation vectors, the model concatenates graph and language vectors, feeds them into the Softmax layer, and then picks the correct answer.

4.4. Experimental Results

4.4.1. Diverse Expansion Methods

In the experiments, we demonstrated the effect of our various expansion methods on pre-trained language models and compared in-depth the ability to answer commonsense questions. To this end, we demonstrated how the AMR graph enables pre-trained language models to understand the semantics of a question and expands the graph with ConceptNet for comprehensive knowledge acquisition. However, expanding the AMR graph with all concepts in the knowledge graph wastes computational resources. Additionally, external knowledge of a few concepts requires a proper reasoning process. We conducted a study to determine which expansion method is the most effective for commonsense reasoning.

For ConceptNet, we expanded the graph based on all words in the question, called the CN full graph. The ConceptNet graph is denoted by $\mathcal{G}_{CN} = \{\mathcal{V}_{cn}, \mathcal{E}_{cn}\}$ and the subgraph of ConceptNet corresponding to the question token is denoted by $\mathcal{G}_{CN}^{token} = \{\mathcal{V}_{cn}^{token}, \mathcal{E}_{cn}^{token}\}$.

The CN full graph (CF) is depicted in Figure 4a and defined as follows:

$$\mathcal{G}_{CF} = (\{\mathcal{V}_{token} \cup \mathcal{V}_{cn}^{token}\} \cup \{\text{root}\}, \{\mathcal{E}_{cn}^{token} \cup \{\text{token}\}\}) \quad (6)$$

The AMR-CN-Full (ACF) graph is an integrated graph in which all nodes of the AMR graph connect to the ConceptNet graph. Additionally, we limited ConceptNet (CN) for the experiments to just two methods. One method was to use the ConceptNet graph corresponding to all question tokens separated by a space in the sentence, as shown in Figure 4b. The graph path learning module could not use the reasoning CN graph owing to the initial disconnection of the question token and disconnection between the concept nodes. Therefore, we combined all tokens from the question to the root node to ensure that our model performed effectively in commonsense reasoning. The ACF graph was identical to the graph before pruning when creating the ACP-ARG.

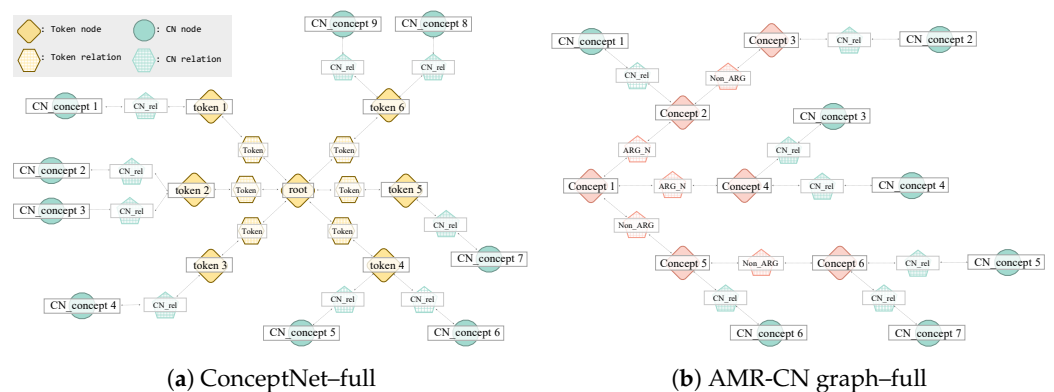


Figure 4. ConceptNet-Full graph (a) expands ConceptNet based on all words in the question. AMR-CN-Full graph (b) expands all the concepts in the AMR graph with ConceptNet graph.

The ACF graph can be expressed as follows: The AMR graph is denoted as $\mathcal{G}_{AMR} = \{\mathcal{V}_{amr}, \mathcal{E}_{amr}\}$. Additionally, the subgraph of ConceptNet matched with the ACF graph is denoted by $\mathcal{G}_{CN}^{AMR} = \{\mathcal{V}_{cn}^{amr}, \mathcal{E}_{cn}^{amr}\}$.

$$\mathcal{G}_{ACF} = (\{\mathcal{V}_{amr} \cup \mathcal{V}_{cn}^{amr}\}, \{\mathcal{E}_{amr} \cup \mathcal{E}_{cn}^{amr}\}) \quad (7)$$

We also conducted ACP-ARG-mini, which is identical to ACP-ARG except for the types of arguments that are pruned. For ACP-ARG, we pruned the nodes that lack edges known as ARG0, ARG1, ARG2, ARG3, and ARG4 with ConceptNet from the ACF graph. Unlike ACP-ARG, ACP-ARG-mini prune nodes lack ARG0 and ARG1 edges, which possess more than 50% of all other argument relations, as shown in Table 1. We expanded the graph based on nodes unrelated to the arguments.

Table 1. Statistics of core roles from AMR graph in CommonsenseQA.

Relation	Ntrain	Ndev	Ntest
ARG0	17,300 (22.70%)	2547 (22.73%)	2477 (23.09%)
ARG1	24,673 (32.38%)	3566 (31.83%)	3521 (32.82%)
ARG2	6001 (7.88%)	864 (7.71%)	829 (7.73%)
ARG3	286 (0.38%)	37 (0.33%)	51 (0.48%)
ARG4	587 (0.77%)	92 (0.82%)	59 (0.55%)
Total relations	76,203	11,204	10,727

The results of the experiments are shown in Table 2. ACP-ARG scored the highest in both the new development and test sets. The performance of the model based on the ACP-ARG graph suggests that using all the information related to the question is not always correct. This suggests that it is efficient and effective when using the specific knowledge graph that the question requires and that the arguments of an AMR graph can provide significant evidence for retrieving the knowledge graph. According to ACP-ARG-mini, the amount of knowledge should be considered even when using arguments from the AMR graph. The model using the ACP-nonARG graph demonstrated inability in the reasoning process.

Table 2. Experiments on diverse graph features.

Language Model	Graph Type	Ndev-Acc(%)	Ntest-Acc(%)	Avg
BERT-base-cased	-	51.81	51.59	52.70
	CN-Full	53.48	53.10	53.29
	AMR-CN-Full(ACF)	53.81	52.38	53.10
	AMR-CN-Pruned-ARG _{0,1} (ACP-ARG-mini)	53.89	52.54	53.22
	AMR-CN-Pruned-nonARG(ACP-nonARG)	53.15	50.77	51.96
	AMR-CN-Pruned-ARG _N (ACP-ARG)	54.38	53.51	53.95

4.4.2. Adversarial Attack Test Using SRL

To demonstrate the analysis of whether the pre-trained language models precisely comprehend the question, we used semantic role labeling (SRL) (https://docs.allennlp.org/models/main/models/structured_prediction/predictors/srl/ (accessed date in 1 April 2021)). SRL [39] labels the predicate and its arguments in a sentence. This study conducted an adversarial attack test on pre-trained language models based on SRL data. For the analysis, we replaced one candidate option, except for the correct answer, with the randomly selected argument related to the predicate of the question in Figure 1a. The experiment assessed whether the model predicted the correct answer using common-sense knowledge or relied on the argument text in the question. We selected BERT as a representative of the pre-trained language models. We tested the model using the CSQA. We fine-tuned the BERT model using the original training dataset for QA tasks and obtained inference results

from the original and SRL-corrupted development datasets. Table 3 shows the results for each development dataset. They indicate that the performance of BERT decreases when one option other than the correct answer is substituted with the argument of the question. Thus, the decrease in performance suggests that BERT merely relies on superficial cues from the question. Our proposed model alleviated the performance of BERT, which showed a decrease of 1.06% compared with the decrease of the fine-tuning model of 5.78%.

Table 3. Adversarial Attack test with Semantic Role Labeling. SRL-C denotes the setting where one distractor is replaced with the random argument in the AMR graph.

Language Model	Setting	Odev-Acc(%)
BERT-base-cased	Original	51.81
BERT-base-cased with ACP-ARG	Original	54.38
BERT-base-cased	SRL-C	46.03 (−5.78%p)
BERT-base-cased with ACP-ARG	SRL-C	53.32 (−1.06%p)

4.4.3. Comparison on Different Language Models

Pre-trained language models based on transformer encoders have been studied since the appearance of BERT [27]. ELECTRA [31] is a model that is trained by replaced token detection using a discriminator. We experimented to determine whether our ACP-ARG graph is effective for diverse pre-trained language models. Additionally, we demonstrated that the proposed graph outperformed the graph representation method of the previous study [34]. Unlike our method that used the model Cai and Lam [32], Lim et al. [34] used the model Guo et al. [16] to generate the AMR graph. Table 4 shows the comparative results based on different types of language models. The input of the language model is “[CLS]+Question+[SEP]+candidate answer”. All language models that used our method outperformed their own fine-tuned score and the other graph reasoning score, achieving 53.95% with BERT and 72.68% with ELECTRA-based models on the average score of our new test set and dev set. The results suggest that the concept representations of the ACP-ARG graph positively affect CSQA and are generalizable to any language encoder.

Table 4. Commonsense QA results on different language encoder types.

Language Model	Ndev-Acc(%)	Ntest-Acc(%)	Avg
BERT-base	51.81	51.59	51.70
ELECTRA-base	71.25	70.19	70.72
BERT-base with ACP-ARG-mini [16,34]	53.97	53.58	53.78
ELECTRA-base with ACP-ARG-mini [16,34]	71.99	70.91	71.45
BERT-base with ACP-ARG	54.38	53.51	53.95
ELECTRA-base with ACP-ARG	73.63	71.72	72.68

4.4.4. Experiment on Official Test Set

Because the ELECTRA-based model with ACP-ARG graphs performed best on the new test set, we evaluated our model using the official training set of ELECTRA-large (1140 examples). For the official test set, the accuracy of our model was 75.79%, as shown in Table 5. The average score was 79.42%, which is higher than that of Lim et al. [34] by 1.37%.

Table 5. Experiment with ELECTRA-large with full training set on the official test set.

Models	Odev-Acc(%)	Otest-Acc(%)	Avg
ELECTRA-large with ACP-ARG-mini [16,34]	82.15	75.43	78.79
ELECTRA-large with ACP-ARG	83.04	75.79	79.42

4.4.5. Experiment on OpenBookQA Dataset

To demonstrate the generalization ability on another reasoning task, we also experimented on OpenBook QA (OBQA). The OBQA dataset consisted of four multiple-choice questions. Elementary-level science-fact-based reasoning is required for this task. As shown in Table 6, the models with the AMR graph or ConceptNet scored 56.00% and 60.00%, respectively, whereas BERT fine-tuning only scored 47.20% and 56.40%. Additionally, ELECTRA-based models outperformed their fine-tuning method, by 64.20% and 82.40% in the official test set.

Table 6. OpenBook QA results on BERT and ELECTRA models.

Language Model	Otest-Acc(%)
BERT-base-cased	47.20
BERT-large-cased	56.40
ELECTRA-base	63.20
ELECTRA-large	77.60
BERT-base-cased with ACP-ARG	56.00
BERT-large-cased with ACP-ARG	60.00
ELECTRA-base with ACP-ARG	64.20
ELECTRA-large with ACP-ARG	82.40

5. Strengths and Limitations

We analyzed the necessity of the semantic representation of the pre-trained language model using adversarial attacks and semantic role labeling. However, we built our data automatically through an external model for the experiments in Table 3. Since humans do not annotate the created data, there may be errors within the data. Additionally, our study suggested a more effective graph representation than the previous study [34]. But, some problems remained. One is the error propagation problem from the AMR construction. The other is the static graph expansion, which might lead the model to learn the same knowledge of different semantic meanings using the same words.

6. Conclusions and Future Works

This paper proposes a strategy of graph representation utilizing the AMR and ConceptNet for commonsense reasoning tasks. The expansion methods of the AMR graph and ConceptNet involved selecting the necessary concepts based on the AMR argument relations because AMR consists of concepts connected with specific logical rules. As a result, extending all nodes connected by argument relations shows the highest performance. However, the proposed method statically expands the same graph information for ambiguous words. Therefore, we plan to use end-to-end common sense inference models, such as AMR constructs and dynamic AMR extension methods that can choose different knowledge for the same word depending on the context of the question.

Author Contributions: Conceptualization, software, investigation, methodology, visualization, writing—review & editing—D.O., J.L.; investigation, visualization, writing—original draft—K.P.; validation, supervision, resources, project administration, and funding acquisition, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques) and the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2022-2018-0-01405) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). In addition, This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2022R1A2C1007616).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: CommonsenseQA: <https://www.tau-nlp.org/commonsenseqa>, accessed date in 1 April 2021, OpenBookQA: <https://allenai.org/data/open-book-qa>, accessed date in 1 April 2021, ConceptNet: <https://conceptnet.io/>, accessed date in 1 April 2021.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Marcus, G. The next decade in ai: Four steps towards robust artificial intelligence. *arXiv* **2020**, arXiv:2002.06177.
2. Berant, J.; Chou, A.; Frostig, R.; Liang, P. Semantic parsing on freebase from question-answer pairs. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 1533–1544.
3. Yih, W.t.; He, X.; Meek, C. Semantic parsing for single-relation question answering. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 22–27 June 2014; pp. 643–648.
4. Yih, S.W.t.; Chang, M.W.; He, X.; Gao, J. Semantic Parsing Via Staged Query Graph Generation: Question Answering with Knowledge Base. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–31 July 2015; pp. 1321–1331.
5. Mitra, A.; Baral, C. Addressing a question answering challenge by combining statistical methods with inductive rule learning and reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 2779–2785.
6. Talmor, A.; Herzig, J.; Lourie, N.; Berant, J. COMMONSENSEQA: A Question Answering Challenge Targeting Commonsense Knowledge. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4149–4158.
7. Mihaylov, T.; Clark, P.; Khot, T.; Sabharwal, A. 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, Brussels, Belgium, 31 October–4 November 2018; pp. 2381–2391.
8. Singh, P.; Lin, T.; Mueller, E.T.; Lim, G.; Perkins, T.; Zhu, W.L. Open mind common sense: Knowledge acquisition from the general public. In Proceedings of the OTM Confederated International Conferences, Rhodes, Greece, 21–25 October 2002; pp. 1223–1237.
9. Lin, B.Y.; Chen, X.; Chen, J.; Ren, X. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. 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), Hong Kong, China, 3–7 November 2019; pp. 2822–2832.
10. Lv, S.; Guo, D.; Xu, J.; Tang, D.; Duan, N.; Gong, M.; Shou, L.; Jiang, D.; Cao, G.; Hu, S. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 8449–8456.
11. Ma, K.; Francis, J.; Lu, Q.; Nyberg, E.; Oltramari, A. Towards generalizable neuro-symbolic systems for commonsense question answering. *arXiv* **2019**, arXiv:1910.14087.
12. Speer, R.; Chin, J.; Havasi, C. Conceptnet 5.5: An open multilingual graph of general knowledge. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
13. Banarescu, L.; Bonial, C.; Cai, S.; Georgescu, M.; Griffitt, K.; Hermjakob, U.; Knight, K.; Koehn, P.; Palmer, M.; Schneider, N. Abstract meaning representation for sembanking. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, Sofia, Bulgaria, 8–9 August 2013; pp. 178–186.
14. Gross, J.L.; Yellen, J.; Zhang, P. *Handbook of Graph Theory*; CRC Press: Boca Raton, FL, USA, 2013; p. 1192.
15. Cai, D.; Lam, W. Graph transformer for graph-to-sequence learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 7464–7471.
16. Guo, Z.; Zhang, Y.; Teng, Z.; Lu, W. Densely connected graph convolutional networks for graph-to-sequence learning. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 297–312. [CrossRef]
17. Vlachos, A. Guided neural language generation for abstractive summarization using Abstract Meaning Representation. *arXiv* **2018**, arXiv:1808.09160.
18. Liao, K.; Lebanoff, L.; Liu, F. Abstract Meaning Representation for Multi-Document Summarization. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 1178–1190.
19. Bonial, C.; Donatelli, L.; Lukin, S.; Tratz, S.; Artstein, R.; Traum, D.; Voss, C. Augmenting Abstract Meaning Representation for Human-Robot Dialogue. In Proceedings of the First International Workshop on Designing Meaning Representations, Florence, Italy, 1 August 2019; pp. 199–210.
20. Issa, F.; Damonte, M.; Cohen, S.B.; Yan, X.; Chang, Y. Abstract meaning representation for paraphrase detection. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 5–6 June 2018; pp. 442–452.
21. Wang, Y.; Liu, S.; Rastegar-Mojarad, M.; Wang, L.; Shen, F.; Liu, F.; Liu, H. Dependency and AMR embeddings for drug-drug interaction extraction from biomedical literature. In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Northbrook, IL, USA, 7–10 August 2017; pp. 36–43.

22. Garg, S.; Galstyan, A.; Hermjakob, U.; Marcu, D. Extracting biomolecular interactions using semantic parsing of biomedical text. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
23. Bonial, C.; Hwang, J.; Bonn, J.; Conger, K.; Babko-Malaya, O.; Palmer, M. English propbank annotation guidelines. In Proceedings of the Center for Computational Language and Education Research Institute of Cognitive Science University of Colorado, Boulder, CO, USA, 14 July 2015; Volume 48.
24. Meyer, C.M.; Gurevych, I. *Wiktionary: A New Rival for Expert-Built Lexicons? Exploring the Possibilities of Collaborative Lexicography*; Linguistics, November 2012. <https://academic.oup.com/book/27204/chapter-abstract/196665268?redirectedFrom=fulltext> (accessed on 1 April 2021).
25. Miller, G.A. WordNet: A lexical database for English. *Proc. Commun. ACM* **1995**, *38*, 39–41. [[CrossRef](#)]
26. Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. Dbpedia: A nucleus for a web of open data. In Proceedings of the Semantic Web, Busan, Korea, 11–15 November 2007; pp. 722–735.
27. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
28. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
29. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 5754–5764.
30. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
31. Clark, K.; Luong, M.T.; Le, Q.V.; Manning, C.D. Electra: Pre-training text encoders as discriminators rather than generators. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
32. Cai, D.; Lam, W. AMR Parsing via Graph-Sequence Iterative Inference. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 1290–1301.
33. Shi, P.; Lin, J. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *arXiv* **2019**, arXiv:1904.05255.
34. Lim, J.; Oh, D.; Jang, Y.; Yang, K.; Lim, H.S. I Know What You Asked: Graph Path Learning using AMR for Commonsense Reasoning. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 2459–2471.
35. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
36. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October; pp. 1724–1734.
37. Shaw, P.; Uszkoreit, J.; Vaswani, A. Self-Attention with Relative Position Representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; pp. 464–468.
38. Salton, G.; Ross, R.; Kelleher, J. Attentive language models. In Proceedings of the Eighth International Joint Conference on Natural Language Processing, Taipei, China, 27 November–1 December 2017; pp. 441–450.
39. Palmer, M.; Gildea, D.; Xue, N. Semantic role labeling. *Proc. Synth. Lect. Hum. Lang. Technol.* **2010**, *3*, 1–103.