Why Do Masked Neural Language Models Still Need Commonsense Repositories to Handle Semantic Variations in Question Answering?

Anonymous RCQA submission

Abstract

Contextualized word representations are now learned by intricate neural network models, such as masked neural language models (MNLMs). The contextualized word representations significantly enhance the performance in automated question answering task which requires to read paragraphs and then extract related phrases. However, identifying the detailed knowledge trained in a MNLM is difficult owing to numerous and intermingled model parameters. This paper provides empirical but insightful analyses on commonsense knowledge included in pretrained MNLMs. First, we propose a test that measures which types of commonsense knowledge could the MNLMs understand. We often observe that MNLMs do not accurately understand the semantic meaning of relations. In addition, based on the difficulty of the question-answering task problems, we observe that the MNLMs are still vulnerable to semantic variations that require commonsense knowledge. We also experimentally demonstrate that we can elevate the performance of existing MNLMs by incorporating information from an external commonsense repository.

1 Introduction

One of long-standing problems in natural language processing (NLP) is to teach machines to effectively understand language and infer knowledge (Winograd 1972). In NLP, reading comprehension (RC) predict the correct answer in the associated context for a given question. RC is widely regarded as an evaluation benchmark for a machine's ability of the natural language understanding and reasoning (Richardson, Burges, and Renshaw 2013).

Neural language models (NLMs) that consist of neural networks to predict a word sequence distribution have widely been utilized in natural language understanding tasks (Radford et al. 2018). In particular, masked neural language models (MNLMs) including BERT (Devlin et al. 2019), that are trained to restore randomly masked sequence of words, have recently led to a breakthrough in various RC tasks. However, the *black box* nature of the neural networks prohibits analyzing which type of knowledge leads to performance enhancement and which type of knowledge remains untrained.

Recently, there are active efforts to understand what is trained on the pretrained NLMs (Conneau et al. 2018; Hewitt and Manning 2019; Tenney et al. 2019; Clark et al. 2019). Existing studies mainly focus on exploring whether a trained model embodies linguistic features for semantic analysis such as tense or named entity recognition (NER), and syntactic analysis such as part-of-speech tagging or dependency parsing for naturally observed texts. One common approach for the linguistic probing is to verify the existence of simple linguistic features by training simple classifiers upon the MNLMs for each task (Conneau et al. 2018).

Commonsense knowledge, defined as 'information that people are supposed to know in common (Nilsson and Nilsson 1998)' and often stored as semantic networks, is known to be another essential factor for natural language understanding and reasoning in the RC task (Mihaylov and Frank 2018). A recent study shows how to attain commonsense knowledge from pretrained MNLMs without additional training procedures (Feldman, Davison, and Rush 2019). However, to the best of our knowledge, detailed analysis on which type of knowledge is trained and untrained in the MNLMs has not yet been thoroughly examined.

Our main focus in this paper is to verify how much the MNLM-based RC models answer or process the complicated RC tasks by understanding semantic relations among the words. To address this, we raise the following questions regarding the semantic understanding of MNLMs:

1. Do MNLMs understand various types of commonsense knowledge, especially relations of attributes? (Section 3.1)
2. Do MNLMs distinguish some semantically related relations well? (Section 3.2)
3. How do MNLM-based RC models solve problems across different levels of difficulty? (Section 4.1)
4. What are the challenging RC task problems for the MNLM-based RC models? (Section 4.2)

To answer Questions 1 and 2, we introduce a *knowledge probing test* designed to analyze whether the MNLMs understand structured semantic commonsense knowledge as semantic triples in an external repository specifically ConceptNet (Speer, Chin, and Havasi 2017). Experimental results on the knowledge probing test reveal that MNLMs understand some types of semantic knowledge. However, unexpectedly, we also observe that MNLMs have a lot of miss-

ing or untrained knowledge, and thus cannot precisely distinguish even some opposite relations.

For Questions 3 and 4, we first define the difficulty of an RC problem with the words overlapped between the context and the question. Then, we analyze how the MNLMs perform on different levels of difficulty and investigate which type of problems be critical limitations of the current MNLMs. As a result of the analyses, we observe that the lexical variation is a crucial determinant in the difficulties of the RC task. In addition, we clarify that the problems that require commonsense knowledge are challenging for the MNLM-based RC models.

Based on the above results, we propose a solution that we can ameliorate the limitations of the current MNLMs by integrating knowledge originated from an external commonsense repository. To verify our solution, we conduct two experiments. Firstly, we manually convert words in the question to integrate the knowledge that is required to solve the problem. Secondly, we propose a neural network architecture that complements MNLMs with the external commonsense repository. In both experiments, we observe that MNLMs could be complemented by integrating commonsense knowledge.

Our main contributions in this paper are as follows:

- From the experimental results of the knowledge probing test on the commonsense knowledge of ConceptNet, we observe that MNLMs have a lot of missing or untrained knowledge.
- By analyzing the results of the MNLM based RC models, we observe that current MNLMs have critical limitations when solving problems requiring commonsense knowledge.
- We empirically verify that MNLMs can be supplemented by integrating external commonsense knowledge manually or automatically.

The paper is organized as follows. Section 2 briefly describes the notions required to readily understand our paper. Section 3 introduces our knowledge probing test and demonstrates the results of the test. Then, we present the performance of the MNLM models on different difficulties of RC problems in Section 4. Section 5 discusses what we observe in the previous sections and propose solutions to ameliorate the limitations. Finally, the conclusion is stated in Section 6. Appendicies can be found in `https://drive.google.com/file/d/1NSeU9Oi-SSQwdfgOeBHysi151K23c-zk/view?usp=sharing`.

2 Background

2.1 Masked Neural Language Models

We consider a MNLM that calculates a probability distribution over the sequence of words with a neural network. Especially, we mainly discuss the BERT and ALBERT (Lan et al. 2019) that are referred to as the MNLMs. Two BERT models[1] (BERT$_{base}$ and BERT$_{large}$) and three ALBERT models[2] (ALBERT$_{base}$ and ALBERT$_{large}$, ALBERT$_{xlarge}$) are used.

BERT and ALBERT models have similar structure made up of the transformer architecture (Vaswani et al. 2017). However, there are two main differences between the BERT and ALBERT. First of all, the ALBERT models (160GB) are trained on much bigger corpus compared with the BERT models (16GB). Furthermore, large and base models of the BERT and ALBERT have same model structure and the ALBERT$_{xlarge}$ has the largest structure among the models that utilized in our experiments. On the other hands, since each ALBERT shares the parameters among the layers, ALBERT models has fewer parameters than the corresponding BERT models.

2.2 Commonsense Repositories

It is important to determine an external resource where we can extract commonsense is necessary. We choose ConceptNet[3], a semantic network widely exploited as a commonsense repository in previous studies (Weissenborn, Kočiskỳ, and Dyer 2017; Wang et al. 2018; Talmor et al. 2019).

ConceptNet, a part of an open mind commonsense (OMCS) (Singh et al. 2002) project, is a semantic network designed to help computers understand the words used by people. ConceptNet includes commonsense knowledge that originates from several resources: crowdsourcing, expert-creating, and games with a purpose.

3 Probing Commonsense Knowledge in MNLMs

This section investigates which types of commonsense knowledge are included in the pretrained MNLMs. Clarifying the trained knowledge is difficult since we deal with the knowledge that has a structured form while the MNLMs have complex and intermingled model parameters. The Cloze test (Chapelle and Abraham 1990), known to be a reliable assessment for the language ability of a participant, is a task wherein one fills in the correct answer for the blank in the text. In the following example, "children and _ are opposite.", the answer word would be 'adults' rather than 'kids'. To infer the correct answer, we must know not only the meaning of each word but also the semantic relation between the words. Inspired by the Cloze test, we introduce a test which we call the *knowledge probing test*.

In the knowledge probing test, we first transform a semantic triple (s, r, o) into a sentence that can be used as an input to a designated MNLM. We generate sentences through predefined predicate templates. For example, a template of the 'Antonym' relation can be "s and o are opposite." For each relation, the template (presented in Appendix A) is selected to be a frequently used pattern representing particular rela-

[1]Uncased models of `https://github.com/google-research/bert`

[2]ALBERT V2 models of `https://github.com/google-research/ALBERT`

[3]ConceptNet 5.6.0, `https://s3.amazonaws.com/conceptnet/downloads/2018/edges/conceptnet-assertions-5.6.0.csv.gz`

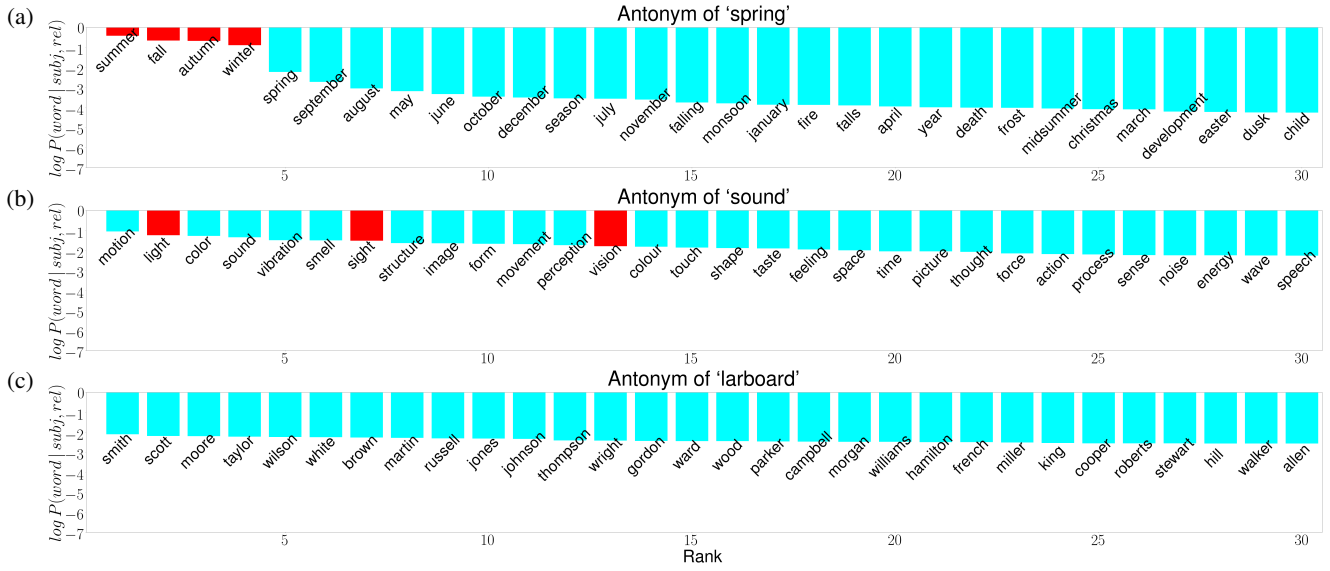

Figure 1: Representative probabilistic distributions of the knowledge probing test results on the BERT$_{base}$ model. (a), (b) and (c) respectively show results of '*antonym* of *sound*', '*antonym* of *spring*' and '*antonym* of *larboard*'. The y-axis indicates $\log_{10}$ probability and the x-axis denotes the ranking of the words. Correct answers are marked in red.

tion in the OMCS dataset.[4]

The object in the generated sentence is masked with a special token '[MASK]' such as "children and [MASK] are opposite." A MNLM then tries to predict the object token [MASK] given a masked sentence. We focus on the objects that comprise a single WordPiece token (Wu et al. 2016) as they are frequently observed in the training corpus. As a result, we can obtain the conditional probability of the masked token and measure the understanding of the MNLMs on the conditional mask-filling task. Our knowledge probing test is similar to recent research (Feldman, Davison, and Rush 2019) in that Cloze test is used. However, the recent study focuses only on the positive results which MNLNs are able to infer semantic knowledge. In contrast, our paper reveals several fundamental limitations of the current MNLMs which are not extensively explored yet due to empirical successes of neural language models.

3.1 Probing on Various Types of Relations

We conduct the knowledge probing test on 37 relations (provided in Appendix A) in ConceptNet to verify whether the MNLMs are properly trained on each relation.

When we visualize the conditional probability of the prediction, we discover that there are three frequently occurred types of distributions. The first type shows a **'L'-shaped** graph, where some words have significantly high probabilities than others. Fig. 1(a) is one example of a 'L'-shaped distribution that shows the probability distribution of the predictions for the antonym of *'spring'*. We can see a drastic drop between the probabilities of 'winter' and 'spring' (the subject of the question), which makes the figure look similar to the character 'L'. The second type shows a half **'U'-shaped** graph, where the probabilities smoothly decrease. Fig. 1(b) is the distribution for the 'sound's antonym', which shows a smooth curve in the distribution. The last type shows a **'–'-shaped** graph, where all candidates share similar probabilities. Fig. 1(c) is the distribution of the 'larboard's antonym', and the graph looks like a bar where no correct answer appears within the top 100 predictions.

Given these empirical observations, we think that semantic triples that show 'L'-shaped graphs are relatively frequently trained on some words as the model is significantly more confident in the words than others. If the semantic triples are properly trained on the model, the words with high probabilities will be the answers as shown in Fig. 1(a). In contrast, we conjecture that the relations are not trained enough in the training when the results show '-'-shaped graphs as the model is not as confident on any of its predictions.

Table 1: Results of micro average and macro average hits@K for the ConceptNet relations.

Model		Hits@K		
		1	10	100
Micro average	BERT$_{base}$	5.93	17.36	34.33
	BERT$_{large}$	5.08	16.78	33.36
	ALBERT$_{base}$	5.15	15.05	31.56
	ALBERT$_{large}$	**8.26**	19.87	35.80
	ALBERT$_{xlarge}$	8.05	**20.34**	**35.94**
Macro average	BERT$_{base}$	5.06	18.84	39.49
	BERT$_{large}$	6.68	20.04	41.94
	ALBERT$_{base}$	4.64	16.54	38.89
	ALBERT$_{large}$	6.18	18.93	39.88
	ALBERT$_{xlarge}$	**7.57**	**22.19**	**42.87**

[4] https://s3.amazonaws.com/conceptnet/downloads/2018/omcs-sentences-more.txt

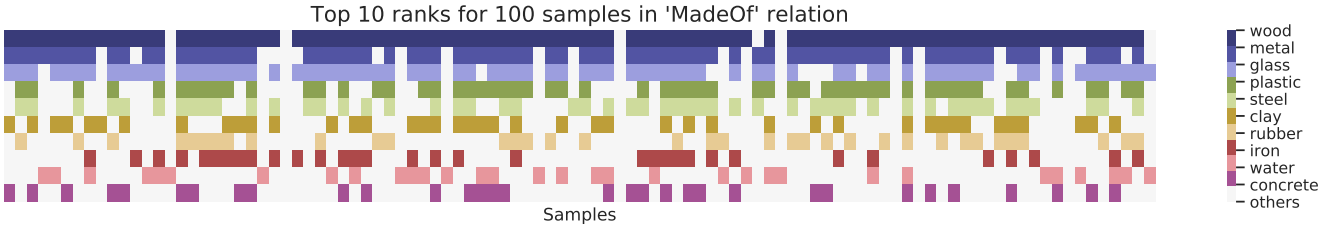

Figure 2: Color-coded results of the BERT$_{base}$ model's predictions on 100 samples in the 'MadeOf' relation. The figure shows whether each sample (x-axis) contains certain object words (y-axis) in the top 10 predictions. Each color represents the 10 most frequently observed words in the predictions on the 'MadeOf' relation.

To quantify the result of the knowledge probing test, we use hits@K metric (Bordes et al. 2013) that measures the ratio of correctly predicted answers, in the top K predictions, out of all true answers from the ConceptNet repository. In the Table 1, we report macro-average, a simple mean of the results of all relations, and micro-average, a weighted mean of the results of the relations according to their frequencies. Individual results on each relation are listed in Appendix B. Large fluctuation can be found in the quantitative results for each relation. Some relations ('DefinedAs', 'IsA', ...) show below 20% in hits@100 while some ('NotCapableOf', 'MadeOf', 'ReceivesAction') show above 70%. The macro-average displays that ALBERT$_{xlarge}$ outperforms other models.

The average hits@100 performances above 30% may seem to be high. However, considering the average number of answers provided by ConceptNet (see Appendix A) is less

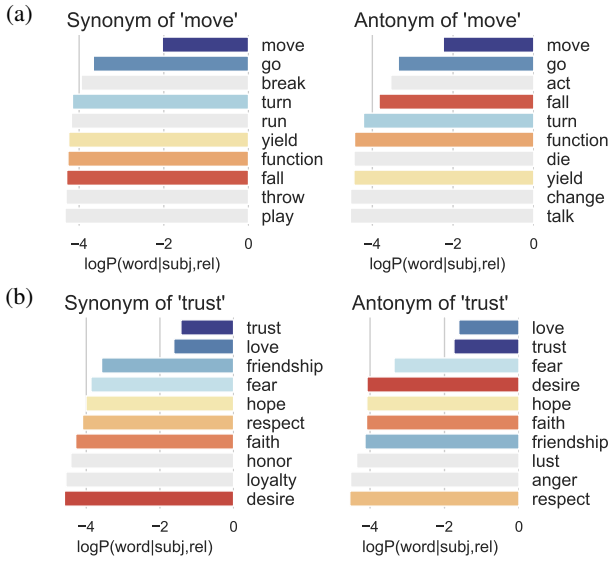

Figure 3: Results of the BERT$_{base}$ model on the top 10 words on the opposite relations on subject words a) 'move' and b) 'trust'. Words commonly observed in both results are painted in the same color, and the other words are in light gray.

than 5, the listed models cannot predict all 5 of the confident answers correctly within top 100 words predicted.

Furthermore, we suspect that the semantic understanding of MNLMs about relations is not as accurate as expected despite the high hit ratios. 'MadeOf' relation is an illustrative example. 'MadeOf' relation shows relatively high performance in hits@10 as around 50% of samples predicted the correct answer within rank 10. However, when we have a closer look at the predictions from MNLMs, it is commonly observed that some specific words are repeated across different subjects. We provide detailed figures in Appendix C.

Especially the BERT$_{base}$ model, which achieves the highest hits@10 in 'MadeOf' relation, presents a noticeable result. Fig. 2 shows the appearance of the 10 most frequent words, in the top 10 predictions of the BERT$_{base}$ model for 100 samples of the 'MadeOf' relation. In more than 70% of sampled subjects, 'wood', 'metal' and 'glass' appear as high-rank predictions. Therefore, our observations say that the prediction tends to follow the marginal distribution of 'MadeOf' relation instead of reflecting the conditional distribution of a subject. This can be problematic when those frequent words are definitely incorrect answers. For example, 'wood' is actually predicted as the most probable answer for the question "What is butter made of?" where the human can easily notice 'wood' is an inadequate answer. As Appendix C, such overlapping of the predict words can be commonly observed among the MNLMs. Note that, as the size of the model increases, the marginality of the 'MadeOf' relation tends to be allayed, but it can be seen that it is not fundamentally solved.

3.2 Probing the Relationship Between Two Opposite Relations

So far, we discuss the behavior of MNLMs for each relation. Here, we address "Do MNLMs precisely understand the semantic difference between relations?" To answer the question, we observe results from the knowledge probing test of four following pairs of opposite relations on the same subject: 'Synonym / Antonym', 'HasProperty / NotHasProperty', 'Desire / NotDesire' and 'CapableOf / NotCapableOf'.

Fig. 3 indicates illustrative examples of the opposite relations on the same subject words. Unexpectedly, there are words simultaneously predicted in the opposite relations.

Table 2: Results of overlapping ratio at top K predictions between the opposite relations.

Model	(Anti/)Relation	Overlap@K		
		1	10	100
BERT$_{base}$	(Ant/Syn)onym	61.19	64.37	68.71
	(Not/)Desires	22.00	57.75	62.52
	(Not/)HasProperty	40.92	46.22	54.01
	(Not/)CapableOf	34.15	50.95	62.71
ALBERT$_{xlarge}$	(Ant/Syn)onym	53.41	58.72	63.07
	(Not/)Desires	63.00	56.25	63.05
	(Not/)HasProperty	31.44	36.03	45.72
	(Not/)CapableOf	43.83	48.13	60.99

Table 3: Experimental results on the incorrect rate between 'Synonym' and 'Antonym' relations.

Model	Relation	Answer	Hits@K	
			10	100
BERT$_{base}$	Synonym	Antonym	30.58	54.16
	Antonym	Synonym	26.25	47.18
ALBERT$_{xlarge}$	Synonym	Antonym	35.45	56.78
	Antonym	Synonym	25.64	47.79

The quantitative results in Table 2 show that words with high probabilities of two opposite relations are common in many cases. The finding supports that the MNLMs may not understand or distinguish the meaning of the opposite relations. This phenomenon is not fundamentally solved as the model grows and the training data increases, as seen in the ALBERT$_{xlarge}$.

To demonstrate that many overlapped words are undesirable, we measure the ratio of incorrect answers by grading the predictions with answers from the opposite relations that are commonly regarded as wrong answers. Among the opposite relation pairs, the answer object of the 'Synonym / Antonym' is incompatible while the others rarely but possibly have the same answer. For this reason, we conduct the experiments on the 'Synonym / Antonym' pair. Hits@K, in this case, can be interpreted as the incorrect rate. As shown in Table 3, the incorrect rate is high even when comparing the ALBERT$_{xlarge}$ with BERT$_{base}$, considering that the no-hit is desirable. In addition, as shown in Appendix B, as the size of the model increases, the performance of each relation enhances, while the incorrect rate also increases as shown in the Appendix D. Thus, we argue that MNLMs with the current training scheme do not discriminate opposite relations well.

4 Analysis on the RC over the Difficulties of the Questions

As reported in the previous section, MNLMs still have incomplete commonsense knowledge. However, MNLM-based RC models outperform other approaches (Radford et al. 2018; Devlin et al. 2019). This section presents results on how MNLMs solve RC questions for different level of difficulties (Section 4.1). Subsequently, we report which types of questions are still challenging for the MNLM-based RC models (Section 4.2).

We analyze BERT$_{base}$, BERT$_{large}$ and ALBERT$_{xlarge}$ models trained on the SQuAD 2.0 RC task dataset (Rajpurkar, Jia, and Liang 2018). This dataset comprises two types of questions: *has answer* and *no answer*. A *has answer* question contains a contextual answer. A *no answer* question does not have a contextual answer. Since we are unable to access the test set of SQuAD, all analyses are conducted with the development set.

4.1 Comparative Studies with Respect to TF-IDF Similarity

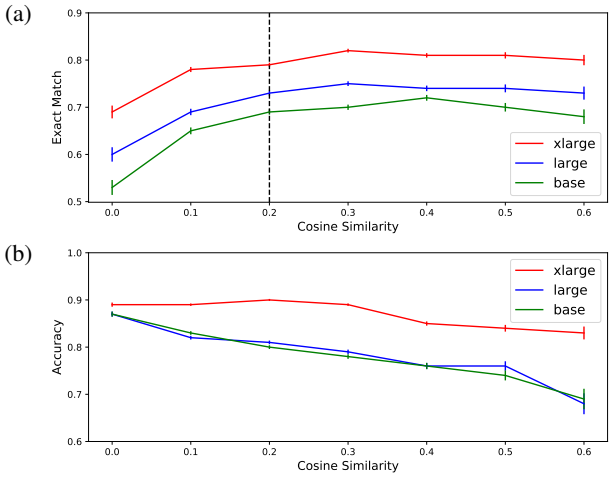

Figure 4: Results on the word overlapping rate and difficulty. X-axis indicates the cosine similarity of context and question, and Y-axis denotes its score. (a) shows results of the *has answer* questions. (b) shows results of the *no answer* questions. Note that, since there is a trivial portion (less than 1%) over similarity 0.6, those questions are ignored.

We look at the difficulty of the RC problem based on a simple hypothesis over lexical overlapping. The hypothesis postulates that the overlap of words in the context and the question strongly correlates with the difficulty level of the RC problem. More specifically, we assume that the *has answer* problem gets easier when the words in the context and question overlap and the *no answer* problem gets harder in a similar situation. To verify our assumption, we analyze the relationship between the lexical overlap of context and question, and the performance of RC models. We calculate the lexical overlap of the context and the question with the cosine similarity between TF-IDF term-weighted uni-gram bag-of-words vectors (Manning, Raghavan, and Schütze 2010). In addition, we set the performance index of the RC task as an exact matching score and an accuracy value for *has answer* and *no answer* questions, respectively.

Fig. 4 shows that the *has answer* questions tend to be more difficult with less lexical overlapping, whereas the *no answer* shows the opposite tendency. In other words, while the model has high performance, the lexical difference between the question and the context still determines the diffi-

Table 4: Examples of easy and hard questions on the *has answer* questions.

Cos. Sim.	Question	Answer	Context
> 0.6	In what year did Savery patent his steam pump?	1698	... In 1698 Thomas Savery patented a steam pump that used steam in direct contact with the water being pumped. ...
< 0.1	Which country was the last to receive the disease?	northwestern Russia	From Italy, the disease spread northwest across Europe, ... Finally it spread to northwestern Russia in 1351. ...

Table 5: Question types and their proportions in each sector. In the models, $Xlarge$, $Large$ and $Base$ indicates ALBERT$_{xlarge}$, BERT$_{large}$ and BERT$_{base}$ respectively. There are 6 question categories and the categories can be tagged with duplicates except for semantic variation and no semantic variation.

Sec.	Models			Question Type						Sampling ratio
				Semantic Variation		Multiple Sentence Reasoning	No Semantic Variation	Others	Typo	
	$Xlarge$	$Large$	$Base$	Synonymy	Commonsense Knowledge					
A	Fail	Fail	Fail	31.32%	**55.49%**	22.53%	16.48%	3.85%	9.34%	182 / 182
B	Pass	Fail	Fail	35.00%	**50.63%**	20.63%	23.13%	1.25%	9.38%	160 / 160
C	Pass	Pass	Fail	**40.19%**	28.97%	17.76%	36.45%	0.93%	4.67%	107 / 107
D	Pass	Pass	Pass	24.53%	10.85%	9.91%	**65.57%**	0.47%	3.30%	212 / 635

culty level of the RC problem. Indeed, the lexical discrepancy between the question and the context requires additional inference to solve, as in the example in Table 4.

4.2 Which Types of Questions Are Still Hard for MNLMs?

We analyze which questions account for the performance differences among the RC models. We begin with dividing the *has answer* questions with less lexical overlapping ($similarity < 0.2$), where relatively difficult questions are classified into four sectors: (A) questions incorrectly answered by all models, (B) questions correctly answered only by the ALBERT$_{xlarge}$, (C) questions correctly answered by the ALBERT$_{xlarge}$ and BERT$_{large}$, and (D) questions correctly answered by all models. For sectors A, B and C, we fully analyze all examples, and for sector D we sample about one-third of the examples (212 out of 635 examples). Then, by referring the question types in (Rajpurkar et al. 2016), we categorize each question into the six classes listed in Table 5. The *synonymy* class means there is a synonym relation between an answer sentence and a question. The *commonsense knowledge* class indicates that commonsense is required to solve a question. The *no semantic variation* category denotes that the question requires neither synonymy nor commonsense knowledge. *Multiple sentence reasoning* class indicates that there are anaphora or clues scattered across multiple sentences. *Others* class indicates that the presented answers are incorrectly tagged. Finally, the *typo* class denotes a typographical error in the question or the answer sentence. Detailed explanations and examples are described in Appendix E. The results show that the proportion of semantic variation-type questions is increased through sector D (easiest) to A (hardiest). Especially, a portion of the commonsense-type questions demonstrate very high in sector A. The results show that it is still challenging for the MNLM-based RC models to deal with the commonsense-type questions.

5 Discussions and Suggested Solutions

Section 3 reveals that, even though MNLMs have the potential to infer commonsense knowledge, there are limitations of the current MNLMs. We conjecture that the current MNLMs are heavily trained to learn 1) the observed information in the corpus and 2) the co-occurrence of the words instead of precise meaning of relations. Furthermore, the results in Section 4 show that MNLM-based RC models have limitations on the semantic variations, more specifically commonsense type questions.

5.1 Can Learned or Learnable External Commonsense Repository Help MNLMs?

As discussed, it is obvious that there are clear limitations in the current MNLM-based RC models and that commonsense knowledge can help ameliorate the limitations. We try to verify whether the external commonsense knowledge can be useful for MNLMs in solving RC problems on the hardest problems in sector A. Details on the experimental settings and examples are provided in Appendix F.

Manual Integration of Commonsense Knowledge First, we introduce a manual approach to incorporate commonsense knowledge from an external repository. Since the difficulty levels of the RC problems are highly affected by the lexical overlap between context and question, we apply a simple paraphrasing rule. For each question, we generate a set of paraphrased questions by replacing a word with its synonym that is in the corresponding context. We use 'Synonym', 'SimilarTo', 'IsA' relations in ConceptNet to find synonyms. We evaluate RC models with the questions in the paraphrased set. For the questions with multiple paraphrases, we choose the best answer for evaluation. We focus on the questions in sector A to see whether it helps to solve the hard questions. As a result, 75 out of 182 questions have a possible paraphrased set and 18 questions are correctly answered by the ALBERT$_{xlarge}$ model. The result implies that

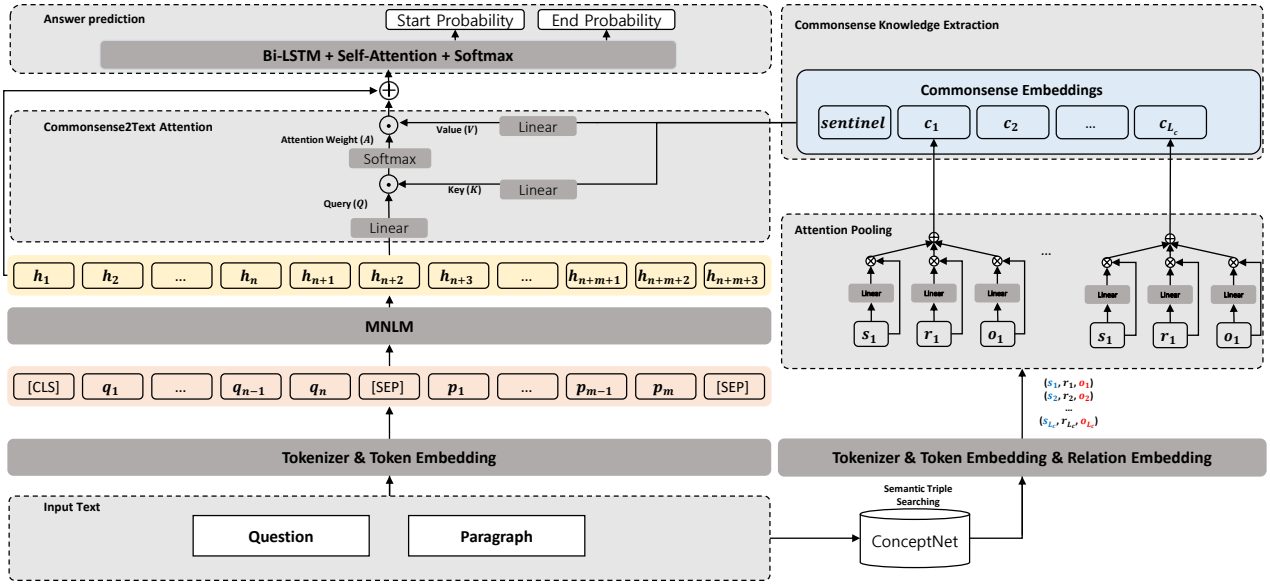

Figure 5: The architecture of our commonsense knowledge incorporated question answering model.

there is a room for performance improvements of MNLMs with the help of external commonsense knowledge. The examples of the manually integrated questions are provided in Appendix F.1.

Automated Integration of Commonsense Knowledge We further design a neural memory network that automatically incorporates the repository to the MNLMs. Fig. 5 shows the overall model comprising four parts: (1) text encoder, (2) commonsense encoder, (3) commonsense2text (C2T), and (4) answer prediction. In the text encoder, the hidden matrix H encodes the question and context through the MNLMs. Then, in the commonsense encoder, we extract commonsense triples whose subject and object are appeared in the text.

Elements of each triple are encoded, then pooled into a single vector through an attention mechanism (Bahdanau, Cho, and Bengio 2014). The triple vectors and a sentinel vector, representing the case where there is no relevant knowledge, are gathered to form a commonsense embedding C. In the commonsense2text, C is selectively fused into H with the following formula, where Q is a linear transformation of H, while K and V are linear transformations of C.

$$A = Softmax(Q \cdot K), I = H + A \cdot V$$

We get an attention weight matrix A, indicating probabilistic weight for V, by adapting a softmax function over the dot product of the Q and V. Then, the result of the Commonsense2Text computed by the dot product of A and V is added to H making a knowledge integrated text matrix I. In the answer prediction, I is input to the bi-directional long short-term memory (Bi-LSTM) then through a self-attention layer and softmax function predicting start and end probabilities of the answer position.

Table 6: Experimental results of the performances when adapting an external commonsense repository. In the table, C2T is an abbreviation of 'commonsense to text' indicating that we integrate the external commonsense repository to the MNLMs.

Model	has ans.		no ans.	overall	
	f1	exact	acc.	f1	exact
BERT$_{base}$	74.31	68.42	79.80	77.06	74.12
+ C2T	**74.96**	**69.32**	**80.64**	**77.80**	**74.99**
BERT$_{large}$	78.33	72.42	80.08	79.21	76.26
+ C2T	**79.41**	**73.84**	**82.98**	**81.19**	**78.41**
ALBERT$_{xlarge}$	**86.00**	**79.52**	88.81	87.41	84.17
+ C2T	84.88	78.83	**90.53**	**87.71**	**84.69**

Table 6 lists experimental results of the MNLMs and our knowledge integrated RC models on SQuAD. The results present that integrating the external commonsense repository yields statistically significant performance improvements of MNLMs. On the other hand, the performance improvements of integrating a C2T to ALBERT $_{xlarge}$ model may seem to be marginal compared to ALBERT $_{xlarge}$. However, we observe that 25 out of the 182 questions in sector A are correctly answered by integrating the C2T module and 22 questions among them are synonymy or commonsense types. The result implies that the C2T module integrated to ALBERT $_{xlarge}$ has potential for helping to solve the limitations of the MNLM-based RC model.

6 Conclusion

In this paper, we investigate which types of commonsense knowledge are trained in the pretrained MNLMs by proposing a knowledge probing test. We find that MNLMs un-

derstand some commonsense knowledge while the trained knowledge is not precise enough to distinguish opposite relations. We also analyze how the MNLM based RC models perform across different difficulty levels of the RC problems and find that questions requiring commonsense knowledge are still challenging to current MNLMs. Finally, we empirically demonstrate that the limitations of MNLMs can be complemented by integrating the commonsense repository.

Acknowledgement

This work is supported by IITP grant funded by the Korea government (MIST) (2017-0-00255, Autonomous digital companion framework and application) and IITP grant funded by the Korea government (MIST) (2017-0-01779, XAI).

References

[Bahdanau, Cho, and Bengio 2014] Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[Bordes et al. 2013] Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of Advances in Neural Information Processing Systems*, 2787–2795.

[Chapelle and Abraham 1990] Chapelle, C. A., and Abraham, R. G. 1990. Cloze method: What difference does it make? *Language Testing* 7(2):121–146.

[Clark et al. 2019] Clark, K.; Khandelwal, U.; Levy, O.; and Manning, C. D. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

[Conneau et al. 2018] Conneau, A.; Kruszewski, G.; Lample, G.; Barrault, L.; and Baroni, M. 2018. What you can cram into a single &!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, volume 1, 2126–2136.

[Devlin et al. 2019] Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

[Feldman, Davison, and Rush 2019] Feldman, J.; Davison, J.; and Rush, A. M. 2019. Commonsense knowledge mining from pretrained models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[Hewitt and Manning 2019] Hewitt, J., and Manning, C. D. 2019. A structural probe for finding syntax in word representations. 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)*, 4129–4138.

[Lan et al. 2019] Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

[Manning, Raghavan, and Schütze 2010] Manning, C.; Raghavan, P.; and Schütze, H. 2010. Introduction to information retrieval. *Natural Language Engineering* 16(1):100–103.

[Mihaylov and Frank 2018] Mihaylov, T., and Frank, A. 2018. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, Volume 1 (Long Papers)*, 821–832.

[Nilsson and Nilsson 1998] Nilsson, N. J., and Nilsson, N. J. 1998. *Artificial intelligence: a new synthesis*. Morgan Kaufmann.

[Radford et al. 2018] Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training.

[Rajpurkar et al. 2016] Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2383–2392.

[Rajpurkar, Jia, and Liang 2018] Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know what you dont know: Unanswerable questions for squad. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, Volume 2 (Short Papers)*, 784–789.

[Richardson, Burges, and Renshaw 2013] Richardson, M.; Burges, C. J.; and Renshaw, E. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 193–203.

[Singh et al. 2002] Singh, P.; Lin, T.; Mueller, E. T.; Lim, G.; Perkins, T.; and Zhu, W. L. 2002. Open mind common sense: Knowledge acquisition from the general public. In *Proceedings of the International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems*, 1223–1237.

[Speer, Chin, and Havasi 2017] Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: an open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4444–4451.

[Talmor et al. 2019] Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4149–4158.

[Tenney et al. 2019] Tenney, I.; Xia, P.; Chen, B.; Wang, A.; Poliak, A.; McCoy, R. T.; Kim, N.; Durme, B. V.; Bowman, S.; Das, D.; and Pavlick, E. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *Proceedings of International Conference on Learning Representations*.

[Vaswani et al. 2017] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems*, 5998–6008.

[Wang et al. 2018] Wang, L.; Sun, M.; Zhao, W.; Shen, K.; and Liu, J. 2018. Yuanfudao at semeval-2018 task 11: Three-way attention and relational knowledge for commonsense machine comprehension. In *Proceedings of the International Workshop on Semantic Evaluation*, 758–762.

[Weissenborn, Kočiskỳ, and Dyer 2017] Weissenborn, D.; Kočiskỳ, T.; and Dyer, C. 2017. Dynamic integration of background knowledge in neural nlu systems. *arXiv preprint arXiv:1706.02596*.

[Winograd 1972] Winograd, T. 1972. Understanding natural language. *Cognitive psychology* 3(1):1–191.

[Wu et al. 2016] Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.