IIT-KGP at COIN - Shared Task: Using pre-trained Language Models for modeling Machine Comprehension

Prakhar Sharma, Sumegh Roychowdhury
Indian Institute of Technology Kharagpur,
India
{prakharsharma, sumegh01}@iitkgp.ac.in

Abstract

In this paper, we describe our system for *COIN 2019 Shared Task 1: Commonsense Inference in Everyday Narrations* Ostermann et al. (2019). We show the power of leveraging state-of-the-art pre-trained language models such as **BERT** (Bidirectional Encoder Representations from Transformers) Devlin et al. (2018) and **XLNet** Yang et al. (2019) over other Commonsense Knowledge Base Resources such as ConceptNet Speer et al. (2018) and NELL Mitchell et al. (2015) for modeling machine comprehension. We used an ensemble of $\text{BERT}_{\text{Large}}$ and $\text{XLNet}_{\text{Large}}$. Experimental results show that our model gives substantial improvements over the baseline and other systems incorporating knowledge bases and got the **2nd position** on the final test set leaderboard with an accuracy of 90.5%.

1 Introduction

Machine Reading Comprehension (MRC) recently has been one of the most explored topics in the field of natural language processing. MRC consists of various sub-tasks Chen et al. (2018), such as cloze-style reading comprehension Hermann et al. (2015); Hill et al. (2015); Cui et al. (2018), span-extraction reading comprehension Rajpurkar et al. (2016) and open-domain reading comprehension Chen et al. (2017), etc. Earlier approaches to machine reading and comprehension have been based on either hand engineered grammars Riloff and Thelens (2000), or information extraction methods of detecting predicate argument triples that can later be queried as a relational database Poon et al. (2010). These methods show effectiveness, but they rely on feature extraction and language tools. Recently, with the advances and huge success of neural networks over traditional feature based models, there have been great interests in building neural architectures for various NLP task Li and Zhou (2018), including several pieces of work on machine comprehension Hermann et al. (2015); Hill et al. (2015); Yin et al. (2016); Kadlec et al. (2016); Cui et al. (2018), which have gained significant performance in machine comprehension domain.

Machine comprehension using commonsense reasoning is required to answer multiple-choice questions based on narrative texts about daily activities of human beings Yuan et al. (2018). The answer to many questions does not appear directly in the text, but requires simple reasoning to achieve. In terms of the nature of the problem, this task can be considered as a binary classification. That is, for each question, the candidate answers are divided into two categories: the correct answers and the wrong answers.

In this paper, we show that pretrained Language Models alone can model commonsense reasoning better than the other models incorporating commonsense knowledge base resources like ConceptNet, NELL, etc integrated with deep neural architectures. We propose to use an ensemble architecture consisting of *BERT* and *XLNet* for this task which achieves an accuracy of 91.0% on the dev set and 90.5% on the test set outperforming the *Attentive Reader* baseline by a large margin of 25.4%.

2 Task Description & Dataset

Formally, this Shared Task: Commonsense Inference in Everyday Narrations Ostermann et al. (2019), organized within COIN 2019 is a multiple-choice machine comprehension task that can be expressed as a quadruple: $< \text{D, Q, A, a} >$ Sheng et al. (2018). Where D represents a narrative text about everyday activities, Q represents a question for the content of the narrative text, A is the candidate answer choice set to the question(this task

Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing, pages 80–84
Hongkong, China, November 3, 2019. ©2019 Association for Computational Linguistics

T	My backyard was looking a little empty, so I decided I would plant something. I went out and bought tree seeds. I found a spot in my yard that looked like it would get enough sunshine. There, I dug a hole for the seeds. Once that was done, I took my watering can and watered the seeds .
Q1	*Why was the tree planted in that spot?*
	to get enough sunshine ✓
	there was no other space ✗
Q2	*What was used to dig the hole?*
	a shovel ✓
	their bare hands ✗
Q3	*Who took the watering can?*
	the grandmother ✗
	the gardener ✓

Figure 1: Example text from SemEval '18 Task 11

contains two candidate answers choice a0 and a1) and a represents the correct answer. The system is expected to select an answer from A that best answers Q according to the evidences in document D or commonsense knowledge.

This task assesses how the inclusion of commonsense knowledge in the form of script knowledge would benefit machine comprehension systems. Script knowledge is defined as the knowledge about everyday activities, i.e. sequences of events describing stereotypical human activities (also called scenarios), for example baking a cake, taking a bus, etc. In addition to what is mentioned in the text, a substantial number of questions require inference using script knowledge about different scenarios, i.e. answering the questions requires knowledge beyond the facts mentioned in the text.

Answers are short and limited to a few words. The texts used in this task cover more than 100 everyday scenarios, hence include a wide variety of human activities. While for question A, it is easy to find the correct answer ("to get enough sunshine") from the text, questions B and C are more complicated to answer. For a person, it is clear that the most plausible answers are "a shovel" and "the gardener", although both are not explicitly mentioned in the texts.

Recently, a number of datasets have been proposed for machine comprehension. One example is MCTest Richardson et al. (2013), a small curated dataset of 660 stories, with 4 multiple choice questions per story. The stories are crowdsourced and not limited to a domain. Answering questions in MCTest requires drawing inferences from multiple sentences from the text passage. Another recently published multiple choice dataset is RACE Lai et al. (2017), which contains more than 28,000 passages and nearly 100,000 questions. The dataset is collected from English examinations in China, which are designed for middle school and high school students.

3 System Overview

We created an ensemble of two systems **BERT** Devlin et al. (2018) and **XLNet** Yang et al. (2019), each of which independently calculates the probabilities of all options for a correct answer.

3.1 Finetuned BERT

BERT is designed to train deep bidirectional representations by jointly conditioning on both left and right context in all layers. We chose $\text{BERT}_{\text{Large, uncased}}$ as our underlying BERT model. It consists of 24-layers, 1024-hidden, 16-heads, and 340M parameters. It was trained on the BookCorpus (800M words) and the English Wikipedia (2,500M words). The context, questions and options were first tokenized with *BertTokenizer* to perform punctuation splitting, lower casing and invalid characters removal. The sequence which is fed into the model is generated in form of [CLS] + context + [SEP] + question + answer + [SEP] for every possible answer and sequence was assigned label 1 for correct answer and 0 otherwise. The maximum sequence length was set as 500 on COIN dataset, with shorter sequences padded and in longer sequences context were truncated to adjust context + question + answer to this length. We first fine-tuned BERT on the RACE dataset using a maximum sequence length of 350 for 2 epochs.

We used the PyTorch implementation of BERT from *transformers*[1] which had the BERT tokenizer, positional embeddings, and pre-trained BERT model. Following the recommendation for fine-tuning in the original BERT approach Devlin et al. (2018), we trained our model with a batch size of 8 for 8 epochs. The dropout probability was set to 0.1 for all layers, and Adam optimizer was used with a learning rate of 1e-5.

[1] `https://github.com/huggingface/transformers`

3.2 Semi-Finetuned XLNet

Both RACE and COIN dataset contains relatively long passages with average sequence length greater than 300. Since the use of the Transformer-XL architecture improves the capability of modeling long sequences besides the AR objective as mentioned in Yang et al. (2019). Hence we focused our attention on XLNet model which is a pre-trained language model built upon the Transformer-XL architecture. We used the XLNet$_{\text{Large, Cased}}$ model which has 24-layer, 1024-hidden and 16-heads. The input to XLNet model is similar to BERT : [A, SEP, B, SEP, CLS], with a small difference that [CLS] token is used at the end instead of the beginning. Here A and B are the two segments, A represents the context and B represents the question + answer. We call our model **semi-finetuned** because we used the *Google Colab TPU* for fine-tuning XLNet but we had to limit the maximum sequence length to 312 owing to the huge computational capacity required by XLNet$_{\text{Large, Cased}}$. We used the Tensorflow implementation of XLNet from *zihangdai/xlnet*[2]. So we couldn't properly fine-tune XLNet on RACE. After fine-tuning on RACE dataset for a few epochs, we fine-tuned further on the COIN dataset keeping maximum sequence length close to 400. The maximum train steps was set to 12000, batch size as 8 and Adam optimizer was used with a learning rate of 1e-5.

3.3 Ensemble

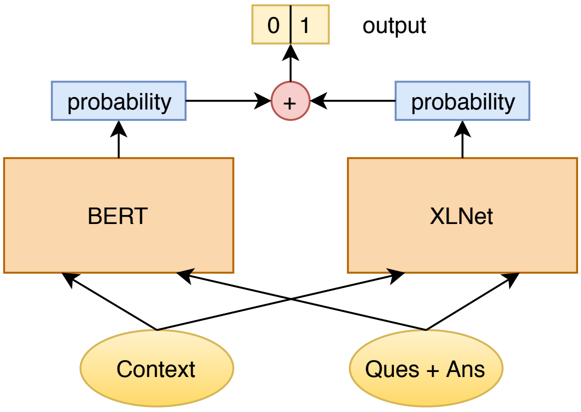

Figure 2: Ensemble Model

Ensemble learning is an effective approach to improve model generalization, and has been used to achieve new state-of-the-art results in a wide range of natural language understanding (NLU) tasks Devlin et al. (2018); Liu et al. (2019b, 2018). For the COIN 2019 shared task, we adopt a simple ensemble approach, that is, **averaging** the softmax outputs from both BERT$_{\text{Large, uncased}}$ and XLNet$_{\text{Large, cased}}$, and make predictions based on these averaged class probabilities. Our final submission follow this ensemble strategy.

4 Additional Experiments

We applied several approaches to the problem that did not generalize as well to the development data and were not included in the final ensemble. Due to space constraints we don't describe the simpler models which include simple rule-based, feature-based classification models, etc.

TriAN: We started with the previous state-of-the-art model for SemEval '18 Task-11 : *TriAN* Wang et al. (2018) as our baseline. We used both SemEval'18 Task 11 and COIN 2019 datasets for training. The Input layer uses GloVe word embeddings concatenated with the part-of-speech tag, named-entity and relation embeddings. It then consists of a Attention Layer which models three way attention between context, question and answer. Question-aware passage representation $\{\mathbf{w}^q_{P_i}\}_{i=1}^{|P|}$ can be calculated as: $\mathbf{w}^q_{P_i} = Att_{seq}(\mathbf{E}^{glove}_{P_i}, \{\mathbf{E}^{glove}_{Q_i}\}_{i=1}^{|Q|})$. Similarly, we can get passage-aware answer representation $\{\mathbf{w}^p_{A_i}\}_{i=1}^{|A|}$ and question-aware answer representation $\{\mathbf{w}^q_{A_i}\}_{i=1}^{|A|}$. These Question-aware passage representation, Passage-aware answer representation and Question-aware answer representation obtained from above are concatenated and fed into 3 BiLSTMs to model the temporal dependency. Then three BiLSTMs are applied to the concatenation of those vectors to model the temporal dependency:

$$\mathrm{h}^q = \mathrm{BiLSTM}(\{\mathbf{w}_{Q_i}\}_{i=1}^{|Q|})$$

$$\mathrm{h}^p = \mathrm{BiLSTM}(\{[\mathbf{w}_{P_i}; \mathbf{w}^q_{P_i}]\}_{i=1}^{|P|})$$

$$\mathrm{h}^a = \mathrm{BiLSTM}(\{[\mathbf{w}_{A_i}; \mathbf{w}^p_{A_i}; \mathbf{w}^q_{A_i}]\}_{i=1}^{|A|})$$

$\mathbf{h}^p, \mathbf{h}^q, \mathbf{h}^a$ are the new representation vectors that incorporates more context information. Then we have question representation $\mathbf{q} = Att_{self}(\{\mathbf{h}^q_i\}_{i=1}^{|Q|})$, answer representation $\mathbf{a} = Att_{self}(\{\mathbf{h}^a_i\}_{i=1}^{|A|})$ and passage representation $\mathbf{p} =$

[2] https://github.com/zihangdai/xlnet

$Att_{seq}(\mathbf{q}, \{\mathbf{h}_i^p\}_{i=1}^{|P|})$. The final output y is based on their bilinear interactions:

$$y = \sigma(\mathbf{p}^T \mathbf{W}_3 \mathbf{a} + \mathbf{q}^T \mathbf{W}_4 \mathbf{a}) \tag{1}$$

Question sequence and answer sequence representation are summarized into fixed-length vectors with self-attention

5 Results and Discussion

This section discusses regarding the results of various approaches we applied in this task. First, as a starting point we ran the best performing model on the SemEval '18 Task 11 - *TriAN* using the same hyper-parameters settings as stated in the paper (Wang et al., 2018). It achieved an accuracy close to 69.0%. We fine-tuned the single BERT$_{\text{Large, uncased}}$ model on the COIN + SemEval '18 Task 11 dataset and that achieved an accuracy of 83.4% on the dev set. We further fine-tuned it on the commonsense dataset RACE for a few epochs which increased the accuracy by 1%. We also fine-tuned the XLNet$_{\text{Large, cased}}$ model on the COIN + SemEval '18 dataset which alone achieved an accuracy of 90.6% on the dev set. But we couldn't fully fine-tune it on the RACE dataset as mentioned earlier in *Section 3.2*. We finally submitted our ensemble system which achieves an accuracy of 91.0% on the dev set and 90.5% on the hidden test set.

We can see there isn't much difference in the accuracy of our final ensemble model on the hidden Test set compared to the Dev set which shows that our model generalizes well to new/unseen data.

Model	**Dev Accuracy**
TriAN	69.0
BERT$_{\text{large, uncased}}$	84.4
XLNet$_{\text{large, cased}}$	90.6
Ensemble	**91.0**

Table 1: Accuracy results for various models.

The main problem with commonsense knowledge bases is that they are hard-coded Wang et al. (2018) and they do not generalize well to hidden dataset. This is also evident from *Table 2* as the top 3 systems (*unofficial leaderboard*)[3] do not use any kind of commonsense knowledge bases.

Rank	**Team**	**Acc.**	**Knowledge Base**
1	PSH-SJTU (Li et al., 2019)	90.6	No
2	**IIT-KGP (ours)**	**90.5**	**No**
3	BLCU-NLP (Liu et al., 2019a)	84.2	No
4	JDA (Da, 2019)	80.7	Yes
5	KARNA (Jain and Singh, 2019)	73.3	Yes

Table 2: Performance comparison among participants of the COIN Shared Task 1, depicting use of commonsense knowledge bases.

6 Conclusion & Future Work

In this paper, we present our system for the *Commonsense Inference in Everyday Narrations Shared Task* at *COIN 2019*. We built upon the recent success of pre-trained language models and apply them for reading comprehension. Our System achieves close to state-of-art performance on this task.

As future work, we will try to explore the in-depth layer by layer analysis of BERT and XLNet attention similar to Clark et al. (2019) and how the attention helps in commonsense reasoning.

Acknowledgments

We would like to thank the **Computer Science and Engineering Department** of **Indian Institute of Technology, Kharagpur** for providing us the computational resources required for performing various experiments. Also, we would like to thank Google for providing us **Google Colab**[4] and free **TPU** (Tensor Processing Unit) access which was extremely crucial for us in fine-tuning XLNet.

References

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions.

Zhipeng Chen, Yiming Cui, Wentao Ma, Shijin Wang, Ting Liu, and Guoping Hu. 2018. Hfl-rc system at semeval-2018 task 11: Hybrid multi-aspects model for commonsense reading comprehension.

[3] https://coinnlp.github.io/task1.html

[4] https://colab.research.google.com/

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does bert look at?an analysis of berts attention.

Yiming Cui, Ting Liu, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2018. Dataset for the first evaluation on chinese machine reading comprehension.

Jeff Da. 2019. Jeff da at coin - shared task.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading childrens books with explicit memory representations.

Yash Jain and Chinmay Singh. 2019. Karna at coin - shared task: Bidirectional encoder representations from transformers with relational knowledge for machine comprehension with common sense.

Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations.

Xiepeng Li, Zhexi Zhang, Wei Zhu, Yuan Ni, Peng Gao, Junchi Yan, and Guotong Xie. 2019. Pingan smart health and sjtu at coin - shared task: Utilizing pre-trained language models and commonsense knowledge in machine reading tasks.

Yongbin Li and Xiaobing Zhou. 2018. Zmu at semeval-2018 task 11: Machine comprehension task using deep learning models.

Chunhua Liu, Shike Wang, , Bohan Li, and Dong Yu. 2019a. Blcu-nlp at coin - shared task: Stagewise fine-tuning bert for commonsense inference in everyday narrations.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019b. Improving multi-task deep neural networks via knowledge distillation for natural language understanding.

Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao. 2018. Stochastic answer networks for machine reading comprehension.

T. Mitchell, W. Cohen, and E. Hruschka. 2015. Never-ending learning.

Simon Ostermann, Sheng Zhang, Michael Roth, and Peter Clark. 2019. Commonsense Inference in Natural Language Processing (COIN) Shared Task Report. In *Proceedings of the 2019 EMNLP Workshop COIN: Commonsense Inference in NLP*.

Hoifung Poon, Janara Christensen, Pedro Domingos, Oren Etzioni, Raphael Hoffmann, Chloe Kiddon, Thomas Lin, Xiao Ling, Alan Ritter, and et al Stefan Schoenmackers. 2010. Machine reading at the university of washington.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. A span-extraction dataset for chinese machine reading comprehension.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text.

Ellen Riloff and Michael Thelens. 2000. A rule-based question answering system for reading comprehension tests.

Yixuan Sheng, Man Lan, and Yuanbin Wu. 2018. Ecnu at semeval-2018 task 11: Using deep learning method to address machine comprehension task.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2018. Conceptnet 5.5: An open multilingual graph of general knowledge.

Liang Wang, Meng Sun, Wei Zhao, Kewei Shen, and Jingming Liu. 2018. Yuanfudao at semeval-2018 task 11: Three-way attention and relational knowledge for commonsense machine comprehension.

Zhilin Yang, Zihang Dai, Yiming Yan, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding.

Wenpeng Yin, Sebastian Ebert, and Hinrich Schutze. 2016. Attention-based convolutional neural network for machine comprehension.

Hang Yuan, Jin Wang, and Xuejie Zhang. 2018. Ynuhpcc at semeval-2018 task 11: Using an attention-based cnn-lstm for machine comprehension using commonsense knowledge.