KARNA at COIN Shared Task 1: Bidirectional Encoder Representations from Transformers with relational knowledge for machine comprehension with common sense

Yash Jain
IIT Kharagpur
Kharagpur, India
yoyo1704@iitkgp.ac.in

Chinmay Singh
IIT Kharagpur
Kharagpur, India
chinmaysingh@iitkgp.ac.in

Abstract

This paper describes our model for COmmonsense INference in Natural Language Processing (COIN) shared task 1: Commonsense Inference in Everyday Narrations. This paper explores the use of Bidirectional Encoder Representations from Transformers(BERT) along with external relational knowledge from ConceptNet to tackle the problem of commonsense inference. The input passage, question and answer are augmented with relational knowledge from ConceptNet. Using this technique we are able to achieve an accuracy of 73.3 % on the official test data.

1 Introduction

Commonsense refers to the skill of making presumptions regarding the physical form, use, behaviour, interaction with other objects etc. that is derived from the naive physics as well as the humans' folk psychology that develops because of the frequent experience that we have as a result of our day to day interaction with these entities.

The task of making commonsense inferences about everyday world is an unsolved and worked upon milestone in the path of Artificial General Intelligence. The approach of attaining this task in the field of Natural Language Processing has seen some advancement in the recent times with the advent of standard Data sets and Tasks like SWAG, Event2Mind and Winograd Schema Challenge.

The general approach followed in natural language processing to judge performance in commonsense inference task is to provide a excerpt of the situation/ event and then some questions are asked relating to the aforementioned paragraph. The model is expected to answer the question which is of the form that cannot be answered by simple extraction of text from the passage but requires certain information that has to be inferred from outside general commonsense resources *i.e.* by the use of commonsense.

Commonsense knowledge is usually exploited by the use of explicit relations (positional, of form etc.) stored in the form of knowledge graphs or binary entity wise relations. Some examples of these databases include Never Ending Language Learner (NELL)(T. Mitchell, 2015), ConceptNet(Liu and Singh, 2004), WebChild(Tandon et al., 2017) etc.

2 Previous Work

Work in development of N.L.P. models that can go beyond simple pattern recognition and use the world knowledge has made progress lately.
Following are some of the major Corpus which have helped make significant progress towards this task:

- **Event2Mind**: (Rashkin et al., 2018) which has 25,000 narrations about everyday activities and situations has been. The best performing model is *ConvNet (Rashkin et al., 2018)*
- **SWAG**: (Zellers et al., 2018) It is a dataset of 113k highly varied grounded situations for commonsense application. BERT Large (Devlin et al., 2018) gives 86.3 percent accuracy on it, which is the current state of the art
- **Winograd and Winograd NLI schema Challenge**: (Mahajan, 2018) Employs Winograd Schema questions that require the resolution of anaphora *i.e.* the model should identify the antecedent of an ambiguous pronoun.

The commonsense information in the form of

Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing, pages 75–79
Hongkong, China, November 3, 2019.

various relations is stored in the form of the following knowledge bases:

- **ConceptNet**: It is a freely-available multilingual language base from crowd sourced resources like Wikitionary and Open Mind Common Sense. It is a knowledge graph with words and phrases as the nodes and relation between them as the edges.

- **WebChild**: It is a large collection of commonsense knowledge, automatically extracted from Web contents. WebChild contains triples that connect nouns with adjectives via fine-grained relations. The arguments of these assertions, nouns and adjectives, are disambiguated by mapping them onto their proper WordNet senses.

- **Never Ending Language Learner**:It is C.M.U.'s learning agent that actively learns relations from the web and keeps expanding it's knowledge base 24/7 since 2010. It has about 80 million facts from the web with varying confidences. It continuously learns facts and also keeps improving it's reading competence and thus learning accuracy.

3 Model

Before getting into the details of our model we first briefly describe the problem statement. Given a scenario, a short context about the narrative texts and several questions about the context, we are required to build a system to solve the question by choosing the correct answer from the choices. We are allowed to use external knowledge to improve our model's common sense inference. For more details, please refer. (Ostermann et al., 2018)
. In our system, we have used BERT(Devlin et al., 2018), a pre-trained representation of unlabelled text conditioned on both right and left sequences. To incorporate commonsense in our model we have used relation knowledge between phrases and words from ConceptNet(Liu and Singh, 2004), a knowledge graph that connects words and phrases of natural language (terms) with labeled, weighted edges (assertions).
Passage, questions and answers were extracted from XML files. Each training example contains a passage $\{P_i\}_{i=1}^{|P|}$, a question $\{Q_i\}_{i=1}^{|Q|}$ and an answer $\{A_i\}_{i=1}^{|A|}$. Each passage is concatenated with

Edge Relation	Event Phrase
$RelatedTo$	A is related to B
$FormOf$	A is a form of B
$PartOf$	A is a part of B
$UsedFor$	A is used for B
$AtLocation$	A is at B
$Causes$	A causes B
$Synonym$	A is synonym of B
$Antonym$	A is antonym of B
$DerivedFrom$	A is derived from B

Table 1: Event Phrases

edge relation from ConceptNet. Method of querying from ConceptNet is inspired from (Wang, 2018), but instead of using a relational vector we convert those relations into event phrases and append them to the passage. The conversion from edge relation to event phrases is given in Table 1. This step is important as edge relations in ConceptNet are not present in vocabulary of pre-trained BERT(Devlin et al., 2018). Event phrases convert the intent of edge relation into words that are present in the vocabulary of pre-trained BERT
Since it is a multiple choice task, every training sample, after augmenting with relational knowledge from ConceptNet is formatted as proposed in (Radford, 2018). Each choice will correspond to a sample on which we run the inference. For a given Swag example, we will create the 4following inputs:
$-[CLS]context[SEP]choice_1[SEP]$
$-[CLS]context[SEP]choice_2[SEP]$
$-[CLS]context[SEP]choice_3[SEP]$
$-[CLS]context[SEP]choice_4[SEP]$
$context$ contains passage concatenated with question and relational knowledge from ConceptNet The model outputs a single value for each input. To get the final decision of the model, we run a softmax over these 4 outputs.

4 Experiments

The training data includes 2500 passages with 14,190 questions while development data has 355 passages and 2019 questions in total. We have used (Pyt) along with Pytorch to read and fine tune pretrained BERT. We have listed the hyperparameters in Table 2. We have tried model and selected the one with best score in development data. We have pretrained the model with Race Dataset (Lai et al., 2017) for 1 epoch. The model is trained on

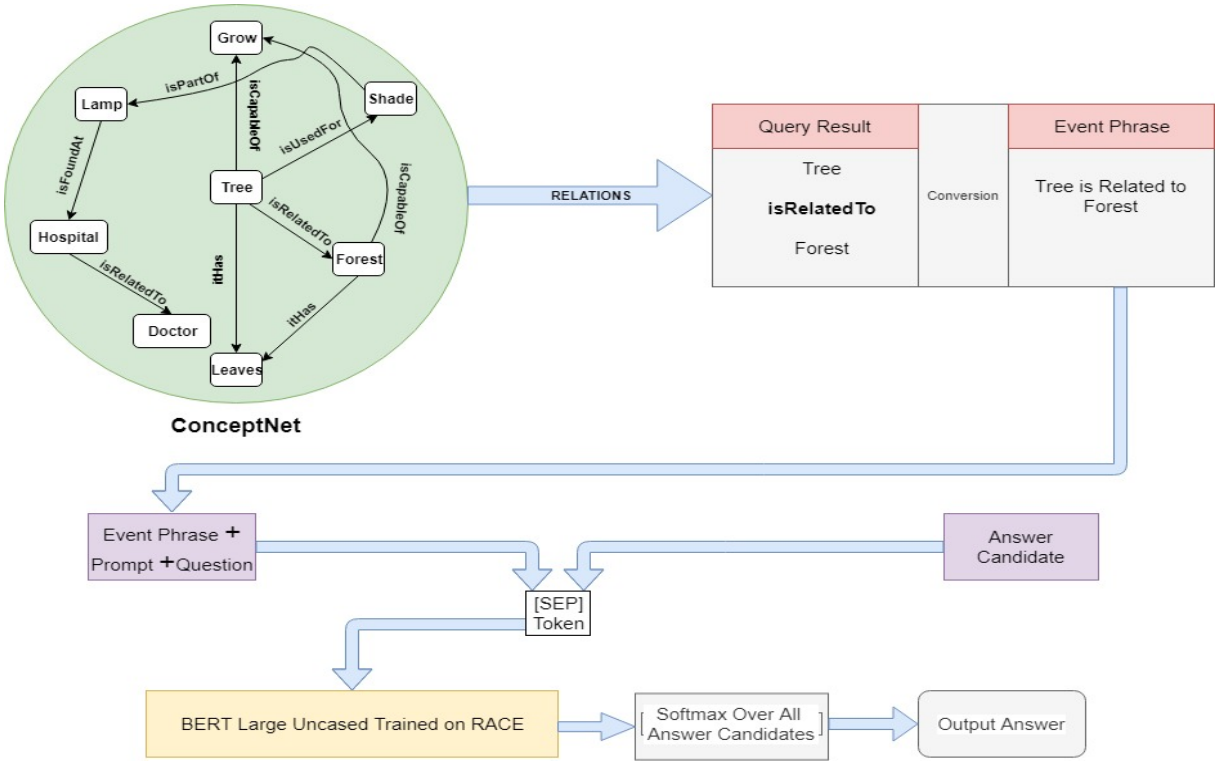

Figure 1: Model Overview

Paramter	Value
$learningrate$	2e-5
$maxseqlength$	210
$batchsize$	4
$epochs$	3
$Optimizer$	ADAM

Table 2: Hyperparameters

Model	Dev-set	Test-set
$w/o\,Q$	**83.6%**	**73.3%**
$w/o\,RACE$	81.2%	-
$w/o\,Q and\,PA\,Rel$	76.7%	-
$w/o\,Q\,and QA\,Rel$	79.8%	-
$w/o\,PA\,Rel\,and QA\,Rel$	74.1%	-

Table 3: Results

Google Colab GPU for 2 epochs. We have used BERT uncased base pretrained model. Gradients are clipped to have a maximum L2 norm of 10.

5 Results

The experimental results are shown in Table 3. The evaluation metric used is accuracy. We have experimented with different variants of $context$. Description of models are given below:

- $w/o\,RACE$: Model without pretraining with RACE and $context$ contains passage, question , relation between passage and answer and relation between question and answer.
- $w/o\,Q$: Model with $context$ containing passage, relation between passage and answer and relation between question and answer.
- $w/o\,Q and\,PA\,Rel$: Model with $context$ containing passage and relation between question and answer.
- $w/o\,Q\,and QA\,Rel$: Model with $context$ containing passage and relation between question and answer.
- $w/o\,PA\,Rel\,and QA\,Rel$: Model with $context$ containing only passage and question.

6 Error Analysis

The reason for difference in accuracy of test set and dev set might be due to the fact that we are using a subset of ConceptNet. The subset was selected based on the vocabulary of training data and development data. The vocabulary

of test data might not be in the selected subset of ConceptNet. There might be few or even no edges for the test data in the selected subset. Thus the accuracy of test data for model $w/o\,Q$ is pretty close to accuracy of dev data for model $w/o\,PA\,Rel\,and Q A\,Rel$.

7 Conclusion

We conculde from our experiments that:

- Pre-Trained Models work better with fine-tuning when the target task for which we are training for is brought into the same domain as the training task. We thus tried with out approach to convert the COIN task as the question anwering task for which BERT was pre trained.

- The addition of ConceptNet derived event phrases increased the model accuracy on the dev set by 9 percent. This is a positive feedback towards the exploitation of the various Knowledge Graphs and Corpora (as mentioned in the introduction). The improvement of accuracy of this method of use of commonsense relations would improve along the the progress of Natural Language Understanding.

- We were not able to use the event phrases on the test set as the edges that we had extracted out of ConceptNet were not inclusive of the test dataset. This problem could be solved if there were enough compute power made available to build and use the whole of ConceptNet or call it from it's web API in the presence of an active internet connection during model evaluation and with sufficient number of call instances of the API available.

8 Scope and Future Work

The Development in Commonsense Inference is detrimental to the progress towards truly general purpose A.I. It's application can be easily be found in development of smarter chat bots and search engines. It delimits the inference systems from using only the provided contextual information from the question asked and hence makes the system more human-like.

Possible developments in this task can come with the use of word embeddings made from ConceptNet and other commonsense corporas and graphs (cite) like Conceptnet Numberbatch embeddings. The accuracy can further be improved by making more grammatically correct and composite sentences from the relations. Further tuning of the Hyperparameters of the model and larger training sample collection would also go long way in helping this field develop.

References:

References

Pytorchtransformer. https://github.com/huggingface/pytorch-transformers.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. pages 785–794.

H. Liu and P. Singh. 2004. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

Vatsal Mahajan. 2018. Winograd schema - knowledge extraction using narrative chains. *CoRR*, abs/1801.02281.

Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018. Semeval-2018 task 11: Machine comprehension using commonsense knowledge. pages 747–757.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2mind: Commonsense inference on events, intents, and reactions. *CoRR*, abs/1805.06939.

Et al. T. Mitchell. 2015. Never-ending learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*.

Niket Tandon, Gerard de Melo, and Gerhard Weikum. 2017. WebChild 2.0 : Fine-grained commonsense knowledge distillation. In *Proceedings of ACL 2017, System Demonstrations*, pages 115–120, Vancouver, Canada. Association for Computational Linguistics.

Liang Wang. 2018. Yuanfudao at semeval-2018 task 11: Three-way attention and relational knowledge for commonsense machine comprehension. *CoRR*, abs/1803.00191.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.