HR@JUST team at SemEval-2020 Task 4: The impact of RoBERTa transformer for evaluation common sense understanding

Heba Al-Jarrah hebaatta96@gmail.com
Rahaf Al-Hammouri rahafalhamouri96@gmail.com
Mohammad AL-Smadi maalsmadi9@just.edu.jo

Department of Computer Science
Jordan University of Science and Technology
Irbid, Jordan

Abstract

This paper describes the results of our team HR@JUST participation at SemEval-2020 Task 4 -Commonsense Validation and Explanation (ComVE) for the POST evaluation period. The provided task consists of three sub-tasks, we participate in task A. We considered a state-of-the-art approach for solving this task by performing RoBERTa model with no Next Sentences Prediction (NSP), dynamic masking, larger training data, and larger batch size. The achieved results show that we got the 11th rank on the final test set leaderboard with an accuracy of 91.3%.

1 Introduction

The concept of common sense in natural language systems has received increasing attention by researchers, as it is considered a long-term problem, and an important challenge in NLP. For example, , *"He drinks milk"*, is a logical sentence, while, *" He drinks apple"*, is not a logic sentence. Looking at the example we know that it is self-evident that a human can distinguish between logical and illogical sentences due to common sense reasoning, while the trained model is weak in common sense and consider as a difficult task in NLP. by using language models trained on large corpora with neural networks (Trinh and Le, 2018; Peters et al., 2018) achieved a good result compared to previous models, but not satisfactory enough. using traditional machine learning algorithm (Ostermann et al., 2018) still have not had good results so far compared with human performance. Although these efforts, it is still unacceptable to rely on them for common sense problems. Natural language understanding is one of the main problems in NLP and the concept of common sense falls under this field. Where it is difficult to distinguish between the meanings of words in terms of semantic meaning and syntax structure. In the SemEval-2020 Task 4 subtask A, the organizers provide a dataset contains of 10 sentences: {s1, s2, o1, o2, o3, r1, r2, r3}. s1 and s2 are two sentences that are similar in syntactic and structure but different in a few words, that require the participants to build a model able to identify which one makes sense.

SemEval is the International Workshop on Semantic Evaluation that has evolved from SemEval.This workshop aims to evaluate semantic analysis systems where is poses some problems related to natural languages (e.g., lexical semantics, Common Sense Knowledge, Knowledge Extraction, Humour, Emphasis, and Sentiment), the SemEval-2020 being the 14th workshop on semantic evaluation. Task 4 (Wang et al., 2020) present three sub-tasks with annotated datasets in the English language. The task for participating teams is to distinguish the sentence that common sense from others. The dataset contains two sentences s1 and s2 are similar in syntactic and structure but different in a few words, that require the participants to build a model able to identify which one makes sense. Further details about Task 4 (Wang et al., 2020) and the dataset in Section 3.2. In this paper, we describe the model was used in our submission in sub-Task A (Validation).

The remainder of this paper is organized as follows: In the second section sheds light on related works that used to solve common sense problems. In the third section, we described the experimental setup, and the methodology proposed in this research. We discussed our results in the section. Finally, Section 5 concludes this research.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 521–526
Barcelona, Spain (Online), December 12, 2020.

2 Related Work

(Wang et al., 2019) they proposed a model to handling the common-sense and explanation problem by using BERT with results 70.1%, 45.61% respectively, on their dataset with 2,021 samples.

Using neural networks with language model produces success in common sense challenging. (Trinh and Le, 2018) proposed an unsupervised approach to handling the commonsense task, they trained and evaluated the Recurrent Neural Network (RNN) model on a large dataset, consist of five corpora namely, SQuAD, CommonCrawl, LM-1-Billion, STORIES, and Gutenberg Books on two levels, character and word level. Their model achieves 70% on the Pronoun Disambiguation dataset and wit accuracy 63.7% on the Winograd Schema dataset.

using seven contextual benchmarks (Zhou et al., 2019) studied and compared unidirectional language models such as GPT, GPT2-medium, and GPT2-base with bidirectional language models, XL Net-base, XLNet-large, BERT-base, BERT-large, RoBERTa-base, and RoBERTa-large for commonsense reasoning tasks. The result shows that the unidirectional models outperform bidirectional models, specially RoBERTa which reaches the best result over the seven contextual benchmarks. but still the human sense the best.

(Chen et al., 2020) proposed a novel approach for sentiment classification by combined multiple data resources for sentiment lexicons on the three-level aspect, sentence, and word. To produce aspect-specific sentence classification the authors used BERT bidirectional model and compare this model with other models like SVM, LSTM, MemNet. The result shows that the BERT is the best model with accuracy = 79.18%.

(Xu et al.,) modified the BERT model and some of the hyperparameter by adding many attention layers after the hidden output of BERT to use for Multiple choice challenge, they evaluated their model using RACE dataset and reach accuracy equal 62.2%, After that, they applied data augmentation and got 66.2%. Finally, they got the best accuracy of 67.9% when using the ensemble model.

(Ostermann et al., 2018) summarized and compared the results of the task from SemEval 2018 on commonsense knowledge for machine comprehensive. 11 teams shared in this competition and used MCSript dataset. The best team achieved accuracy equal to 83.95% using LSTM as the main model, but still this accuracy so far from the human (98%).

(Rajani et al., 2019) collected dataset for commonsense reasoning called Common Sense Explanation (COS-E) from human explanations. Then they used this dataset to train the model to automatically produce explanations and called this novel Common-Sense Explanation (CAGE). They used the BERT model as a baseline with commonsense question answering input and achieved accuracy equal to 63.8. when they used inputs with COE-S, the accuracy improved to 65.5. Finally, the CAGE improved the accuracy to 72.6%.

(Talmor et al., 2018) created the COMMONSENSEQA dataset for commonsense question answering tasks, which contains 12247 questions. They used this dataset to train and evaluate several models such as GPT, BERT-large, and ELMO. BERT-large achieved the best accuracy of 55.9%, but this result is still little compared to the human accuracy 98%.

(Chen et al., 2018) proposed Hybrid Multi-Aspects (HMA) model, that combined two techniques for multiple choice task, which compute attention and similarity for text, choices, and question then integrated the results with RNN model prediction. They achieved accuracy = 84.13%.

(Yuan et al., 2018) proposed attention-based CNN-LSTM model for commonsense knowledge. They used attention technique and combine CNN and LSTM models by convert the text to vector using embedding layer then fed into CNN and finally, the output feature from CNN layer fed into LSTM. Their model accuracy = 71.43%.

(Han et al., 2013) proposed three systems for the semantic similarity between pairs of sentences, they trained and evaluated their system using a large dataset collected from many resources. They used a combination from part-of-speech tagging, wordNet, and LSA for lexical features. The first proposed system used term alignment and achieved the highest score 89, the other systems used support vector regression.

Semantic text similarity has several metrics and we can use several deep learning models to evaluate it. (Ramaprabha et al., 2018) presented a survey that described these metrics and models. As well as, they

ID	Sent0	Sent1	Label
1951	He wrote an exam in pen	He wrote an exam in knife	1
1190	Dragonfly can fly	shoes can fly	1
1785	He put a yacht in bed.	He put the pillow on the bed.	0

Table 1: Dataset Examples.

discussed the challenges and the evaluation applications that they found.

3 Methodology

3.1 Task Description

We will provide a brief introduction of the SemEval-2020 Task 4 Commonsense Validation and Explanation (ComVE)(Wang et al., 2020) and then we will display at a glance an of the RoBERTa model that we used to solve the task in Figure 1. The organizer for this task, have provided three sub-task namely: Task A; Validation, Task B; Explanation (Multi-Choice), and Task C; Explanation (Generation) for English language which this task inspired by (Wang et al., 2019). In this paper, we participated in Task A (Validation) which contains two similar statements in the syntactic structure but with some differences in a few words, the participants are required this subtask to build a model able to distinguishes which sentence makes sense or not.

3.2 Dataset

In this project, a dataset provided by SemEval2020 task4 (Wang et al., 2020) completion is involved. This dataset provided for Commonsense Validation and Explanation task including three subtasks, our focus is in Sub-Task A, the dataset for all sub-tasks available on Github at

. Dataset consists of four files, trail which contains 2021 rows, train which contains 10000 rows, development which contains 997 rows, and finally test which contains 1000 rows, each file contains two sentences one of them is against common sense. Furthermore, the dataset available with the gold labels for all files except test files. Table 1 provides a set of examples for Task A. For data pre processing, we remove all punctuations and all words convert to lower-cased.

3.3 Proposed System

BERT (Devlin et al., 2018) is a bidirectional transformer released by google in 2018 is a new state-of-art technique in NLP, helped to solve several NLP problems such as Common sense, Question answering, and SWAG tasks. BERT solve the context learning limitation problem that found in directional models using Next Structure Prediction(NSP) and Masked Language Model (MLM). Where MLM can predict 15% from tokens. BERT based on attention techniques to learn the contextual relationships between words in a text. The BERT trained using 3.3 billion tokens. Bert has two versions, BERT-base and BERT-large, which differ in the number of layers, parameters, and attention heads.

XLNet (Yang et al., 2019) is a large bidirectional transformer that uses permutation language modeling. XLNet outperforms BERT in 20 different tasks like document ranking, sentiment analysis, and natural language inference. XLNet predicts all tokens in random order, while BERT only predicts 15% tokens. In Addition, XLNet uses XL transformer in the architecture. XLNet has two versions, XLNet-base and XLNet -large. XLNet-base trained on the same BERT dataset with 110 million parameters, while XLNet-large trained on 33 billion tokens taken from English Wikipedia, Giga5, Common Crawl, BookCorpus, and ClueWeb with 340 million parameters.

https://github.com/wangcunxiang/SemEval2020-Task4-Commonsense-Validation-and-Explanation Commonsense Validation and Explanation Dataset

	Methods	Parameters(Million)	Dataset size
BERT	Bidirectional Transformer with MLM and NSP	BERT-base: 110, BERT-large: 340	16GB
RoBERTa	Bidirectional Transformer without NSP	RoBERTa-base: 110, RoBERTa-large: 340	160GB
XLNet	bidirectional transformer with permutation language modeling	XLNet-base: 110, XLNet-large: 340	XLNet-base: 16GB, XLNet-large:113GB

Table 2: Models comparison.

parameter	Value
number of epochs	6
batch size	16
optimizer	adam_epsilon
learning rate	4e-6

Table 3: Proposed Model Hyper-parameters

RoBERTa (Liu et al., 2019) (Robustly optimized BERT approach) proposed by Facebook. RoBERTa has the same BERT architecture but re-trained using a large amount of data, larger batch size, and without Next Sentence Prediction (NSP). RoBERTa uses dynamic masking which in each epoch the masked token is changed. The train dataset size equal 160GB collected from OpenWebText, CCNEWS, BookCorpus, and STORIES, 16GB from this data is the BERT dataset. RoBERTa has two versions, RoBERTa-base and RoBERTa -large. RoBERTa -base has 110 million parameters, while RoBERTa -large has 340 million parameters. **Figure 1** presents model architecture. The mechanism of the work of RoBERTa model is very similar to the BERT model, but with some differences that make the RoBERTa model different from the BERT, which is represented with no Next Sentences Prediction (NSP), dynamic masking, larger training data, and larger batch size. As we see in figure 1. the model receives two sentences together, as input data, which uses some reserved words like $\langle s \rangle and \langle /s \rangle$,to indicate the starting and end of sentences and to distinguish between two sentences ,then a positional embedding comes to give each tokens specific positions in a sequence, then RoBERTa model makes a prediction (0 and 1) where one indicated for sentences that not make a sense and zero indicated for sentences make a sense. Table 2 illustrates the models comparison.

4 Results

To evaluate the performance of the performing models accuracy has used according to the competition instruction. In order to evaluate our developed RoBERTa model, we used PyTorch for implementation. We optimized the hyperparameters during the training step, Table 3 shows the best values of the parameters. We train the developed model for 6 epochs with a learning rate of 4e-6 and batch size 16, however, *adam_epsilon* was used.

Our participation in this task presented the results for the POST evaluation. We have used several models during the experimental step, for instance, RoBERTa, BERT, XLnet with based version, USE, and Elmo. We used ELMO model as baseline for our experiments, furthermore, RoBERTa outperforms all the performed models. As shown in Table 4 for all models, the best performed model achieved an accuracy= 91.3% using pseudo label technique (Lee, 2013). RoBERTa demonstrated the BERT model with an enhancement of 4.9%.

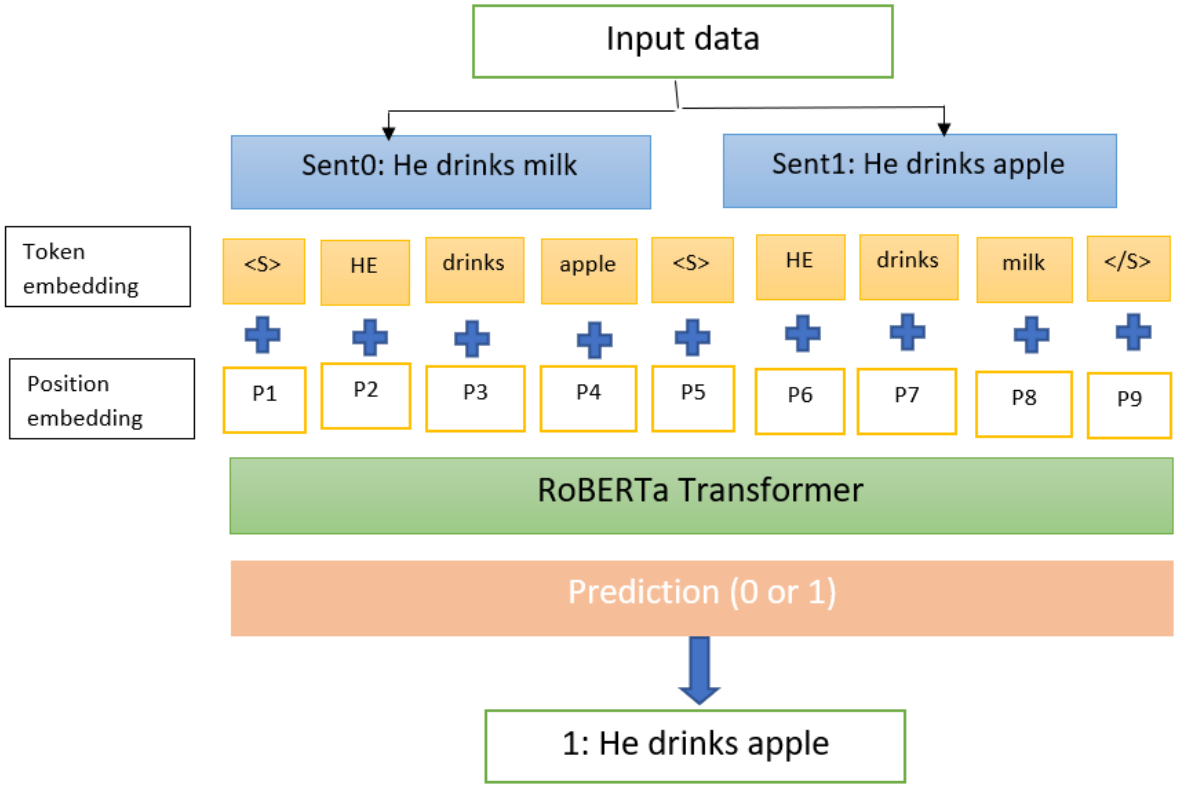

Figure 1: The architecture of our model

Model	Accuracy
RoBERTa (Liu et al., 2019)	90.3%
RoBERTa + pseudo label teq (Liu et al., 2019)	**91.3%**
BERT (Devlin et al., 2018)	86.4%
XLnet-base (Yang et al., 2019)	84.2%
USE (Cer et al., 2018)	72%
Elmo (Peters et al., 2018)	64.9%

Table 4: Experimental Results.

5 Conclusion

In this paper, we have presented our participation at SemEval-2020 Task 4 - Commonsense Validation and Explanation (ComVE). Several models have presented in particular transformers to tackle Task A (Validation) that aims to differentiate between two sentences against common sense among them. RoBERTa was used as the main model to solve this task. The achieved results show that we got the 11th rank on the final test set leaderboard with an accuracy of 91.3%. For future work, we will try to study the impact of transfer learning.

Acknowledgements

This research is partially funded by Jordan University of Science and Technology.

References

Daniel Matthew Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *ArXiv*, abs/1803.11175.

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. *arXiv preprint arXiv:1803.05655*.

Fang Chen, Zhigang Yuan, and Yongfeng Huang. 2020. Multi-source data fusion for aspect-level sentiment classification. *Knowledge-Based Systems*, 187:104831.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Lushan Han, Abhay L Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. 2013. Umbc_ebiquity-core: Semantic textual similarity systems. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 44–52.

Dong-Hyun Lee. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 2.

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

Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018. Semeval-2018 task 11: Machine comprehension using commonsense knowledge. In *Proceedings of the 12th International Workshop on semantic evaluation*, pages 747–757.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*.

J Ramaprabha, Sayan Das, and Pronay Mukerjee. 2018. Survey on sentence similarity evaluation using deep learning. In *Journal of Physics: Conference Series*, volume 1000, page 012070. IOP Publishing.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019. Does it make sense? and why? a pilot study for sense making and explanation. *arXiv preprint arXiv:1906.00363*.

Cunxiang Wang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Yue Zhang. 2020. SemEval-2020 task 4: Commonsense validation and explanation. In *Proceedings of The 14th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Kegang Xu, Jingjie Tin, and Jungyoun Kim. A bert based model for multiple-choice reading comprehension. *Passages*, 6(368):362.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Hang Yuan, Jin Wang, and Xuejie Zhang. 2018. Ynu-hpcc at semeval-2018 task 11: Using an attention-based cnn-lstm for machine comprehension using commonsense knowledge. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1058–1062.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2019. Evaluating commonsense in pre-trained language models. *arXiv preprint arXiv:1911.11931*.