NLP@JUST at SemEval-2020 Task 4: Ensemble Technique for BERT and Roberta to Evaluate Commonsense Validation

Emran Al Bashabsheh
Computer Science
Jordan University of
Science and Technology
Irbid, Jordan
emranalbashabsheh@gmail.com

Ayah Abu Aqouleh
Computer Science
Jordan University of
Science and Technology
Irbid, Jordan
ayaalkhader96@gmail.com

Mohammad AL-Smadi
Computer Science
Jordan University of
Science and Technology
Irbid, Jordan
masmadi@just.edu.jo

Abstract

This paper presents the work of the NLP@JUST team at SemEval-2020 Task 4 competition that related to commonsense validation and explanation (ComVE) task. The team participates in sub-taskA (Validation) which related to validation that checks if the text is against common sense or not. Several models have trained (*i.e. Bert, XLNet, and Roberta*), however, the main models used are the RoBERTa-large and BERT Whole word masking. As well as, we utilized the results from both models to generate final prediction by using the average Ensemble technique, that used to improve the overall performance. The evaluation result shows that the implemented model achieved an accuracy of 93.9% obtained and published at the post-evaluation result on the leaderboard.

1 Introduction

Recently, Natural Language Processing (NLP) has got more attention because it is able to deal with human and machine communication without human intervention (Lytinen, 2005). NLP is a subset of Artificial Intelligence (AI) that grants the machines ability to read and understand human words using Machine Learning (ML) and Deep Learning (DL) models. However, it is used to solve a large scope of tasks such as Machine Translation, Speech Recognition, Sentiment Analysis, etc. NLP has two subsections Natural Language Understanding (NLU) and Natural Language Generation (NLG). Furthermore, NLU with commonsense has gained the researcher's interests in the last few years(Wang et al., 2019).

Commonsense is known as the knowledge that captured about the world's continents and the implicit reasoning process (Storks et al., 2019). As known the human have obtained commonsense knowledge from their daily life, they can distinguish the sentence is against commonsense or not from their previous knowledge. The Commonsense task is considered as one of the important tasks of NLU in NLP that makes the machines more intelligent (Storks et al., 2019). Hence, there are many works in the literature focus on commonsense tasks, and various benchmarks created to evaluate a trained model's capabilities to capture the commonsense validation task. For example, (Wang et al., 2019) introduced a new benchmark that has two tasks which are: the Sen-Making task and Explanation task, to evaluate the system can judge the text makes sense or not and provide reasoning why the text is not commonsense (Wang et al., 2019).

This paper presenting NLP@JUST teamwork submitted at SemEval-2020 Task 4 competition at a post-evaluation time. This task is related to commonsense validation and reasoning, released by (Wang et al., 2020). SemEval-2020 Task 4 has three subtasks: Validation, Explanation (Multi-Choice), and Explanation (Generation). Specifically, this paper focuses on the commonsense validation task. Mainly the RoBERTa (Liu et al., 2019) and BERT-WWM (Whole Word Masking)(Devlin et al., 2018) models are adopted, where they are considered as an extension to the BERT (Devlin et al., 2018) model. The work obtained good results with 93.9% accuracy by using the average ensemble for both of models, which is perfect when compared with other pre-trained models we used (*i.e. RoBERTa (Liu et al., 2019), BERT original (Devlin et al., 2018), XLNet (Yang et al., 2019), ELMO (Peters et al., 2018), USE (Cer et al., 2018)*).

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 574–579
Barcelona, Spain (Online), December 12, 2020.

The rest of the paper is organized as follows: Section 2 provides the most relevant works, Section 3 describes the details of the methodology and proposed system, Section 4 presents the experiments that were conducted in addition to the results obtained, and finally, Section 5 concludes the work.

2 Related Works

There is a growing attention towards common sense in the research community and many previous studies tried to build and use machine learning and deep learning techniques for dealing with common sense tasks. For example, (Trinh and Le, 2018) introduced a method to apply language models on tasks that deal with a specific of commonsense knowledge. They adapt the probabilities that computed by the language models to evaluate the statement if it is true or not.

The most recent pre-trained model called RoBERTa (Liu et al., 2019) which known as A Robustly Optimized BERT Pretraining Approach. They developed the previous BERT model by training the model longer time, and increasing batches using more data in addition to removing the objective of next sentence prediction. Besides that, they trained the model on longer sequences, and changing the masking pattern was used in the training phase.(Zhou et al., 2019) used the Uni-directional language modeling (GPT, GPT2-base, GPT2-medium) and Bi-directional language modeling (Bert-base, Bert-large, XLNet-base, XLNet-large, Roberta-base, and RoBERT-alarge) to compare them on seven benchmarks realted to commonsense reasoning tasks. The results showed that Roberta outperformed other pre-trained models over the seven benchmarks.

(Wang et al., 2019) produced a new benchmark for evaluating the system that abilities to capture the sentence that do not make sense from the sentence that makes sense. Besides, they can explain why this sentence does not make sense. Several models used on their experiments which are: Random, ELMO, BERT, and fine-tuned ELMo, and (Ostermann et al., 2018) developed a framework that provided an evaluation for commonsense knowledge for the set of machine comprehension as a subtask of SemEval 2018. The best-ranked model achieved by (Chen et al., 2018) proposed a neural network model called Hybrid MultiAspects (HMA). Their model aim was to provide a multi-aspect output and then combine it along together for the output of the final prediction. They achieved the best rank on this SemEval task with 84.13% accuracy. Also, for the same task, (Wang et al., 2018) applied a three-way attention technique for interaction among texts, questions, and answers on the top of BiLSTMs. However, they have achieved an accuracy of 83.95%.

(Chen et al., 2020) introduced a novel approach for sentence classification by combining the corpora on the three-level aspect, sentence, and word sentiment lexicons. As well as they employed the BERT to produce an aspect-specific sentence classification.

3 Methodology

3.1 Task Description

SemEval 2020 (Wang et al., 2020) has published Task 4: Commonsense Validation and Explanation. This task inspired by previous work proposed by (Wang et al., 2019). Task4 consists of three subtasks A, B, and C for Validation, Explanation (Multi-Choice), and Explanation (Generation) respectively. Generally, the purpose of task 4 is to recognize the sentences that achieve makes-sense from don't achieve a logical sentence. In this research, we focused on sub-taskA that consists of two sentences correspond to each other; where the first sentences achieve the make-sense and the other against make-sense.

3.2 Dataset

Commonsense validation and explanation dataset proposed by (Wang et al., 2020) is being used to evaluate the implemented model,and it publicly available [1]. The provided dataset is created according to the inspiration of annotators themselves depending on raw English sentences of ConceptNet5.5 (Speer et al., 2017). Three annotators have examined the dataset case by case. Each sub-task has its own human evaluation, for instance, sub-taskA three annotators have to answer each case the same evaluation.

[1] https://github.com/wangcunxiang/SemEval2020-Task4-Commonsense-Validation-and-Explanation, Commonsense Validation and Explanation dataset

However, they should rewrite or abolish the sentence. This dataset is used to evaluate three sub-tasks, we focus in this work on sub-taskA. The Dataset consists of three main files namely: train; which includes 10000 rows for each sentence and their labels, development; which includes 997 rows, and test includes 1000 rows including trail data with 2021 rows that aim to obtain the text is against common sense. Table 1 provides examples of the provided dataset.

id	sent0	sent1	label
1	He drinks apple.	He drinks milk.	0
4	A niece is a person.	A giraffe is a person	1
422	Students teach teachers	Teachers teach students	0

Table 1: Examples are used to evaluate sub-taskA

3.3 Dataset pre-processing

In this stage, some pre-processing techniques were applied, as hoped to increase the accuracy, but the attempts have proved the opposite. Many steps for pre-processing have experimented and many procedures were applied, Table 2 shows examples for these processes. More illustrations of what pre-processing steps used are listed as follows::

1. Removing punctuations: symbols such as exclamation mark, backslash, comma, etc.
2. Removing identifiers: identifiers and linking words in a language are removed.
3. Removing stopwords: where the most commonly used words in a language are removed, as they usually do not bring additional meaning.
4. Expanding abbreviations: where shortcuts for terms are expanded to their representative words.
5. Lemmatization: This reduces the inflection of the words properly to ensure that the root of the word belongs to the language.

Technique	Examples
Removing punctuations	”!, +, :, ;, ?, @”
Removing identifiers	”the”, ”a” and ”an”
Removing stopwords	”He”, ”They”, ”is” and ”on”
Expanding abbreviations	”I’m”, ”can’t” into ”I am”, ”Can not”
Lemmatization	”been”,”had” into ”be”,”has/have”

Table 2: Examples of Pre-processing Steps

3.4 Proposed System

The Transformer network structure proposed by (Vaswani et al., 2017), the major intention is to tackle sequence-to-sequence NLP tasks which trying to deal with long-range dependencies easily. This technique presented the encoder-decoder structure based on attention layers which become the solution for the state of art NPL tasks. The transformer structure allows the ability of the input sequence to passed in parallel, which aims to utilize the GPU efficiently leading to an increase in the speed of the training process. Google used a transformer in the proposed pre-trained model BERT (Devlin et al., 2018). Furthermore, Facebook proposed an enhanced version of BERT called RoBERTa.

The robustly optimized BERT approach (RoBERTa) (Liu et al., 2019) is one of the models that has enhanced the capability of the BERT model and improved the performance over the domain

of variety benchmarks. It has a few differences in the training mechanism and the inner composition from Bert model, where Bert has two objectives Masked Language Model (MLM) and Next Sentence Prediction (NSP). Masked Language Model (MLM) is a technique that selects 15% of the tokens randomly and replaces with a certain token [MASK]. Besides, Bert deals with masks as a static mask from the beginning of the training to the end. In the second objective, Next Sentence Prediction (NSP) has been excluded from Roberta. Roberta has a dynamic mask that exchanges the tokens each training iteration. Furthermore, Roberta has been trained on a longer sequence of tokens than Bert and the model trained more extra times with a bigger batches size. Roberta has been developed through pre-trained cumulatively improvement from 16GB until 160GB of text data, as well as, the model pre-trained more extra times over 100K, 300K and 500K of batch size steps.

On the other hand, we utilized Bert whole word masking as it is a newly released version of Bert. Bert whole word masking (WWM) comes to address the masking partial subword tokens. For example, the statement 'the in ##vert ##er was able to power the continent' is an instance from data that trained in Bert-original and Bert-WWM: Input Text: 'the in ##vert ##er was able to power the continent'. Original Masked Input: 'the in [MASK] ##er was able to [MASK] the continent'. Whole Word Masked Input: 'the [MASK] [MASK] [MASK] was able to power the continent'. In the example above shows, the phrase 'inverter' is tokenized into three word pieces encoding 'in ##vert ##er', which the Bert-WWM is taken into account masking subwords token that consists of the word encoded.

In this paper, we adopted the Roberta and Bert-WWM models as the major models by using the average ensemble technique to solve the provided sub-task that aims to recognize which of the input sentences are against commonsense or not. we produced that technique by taking the average prediction of the results from a couple of pre-trained models and use it to obtain a final prediction. As depicted in figure 1 that shows the implemented ensembling technique.

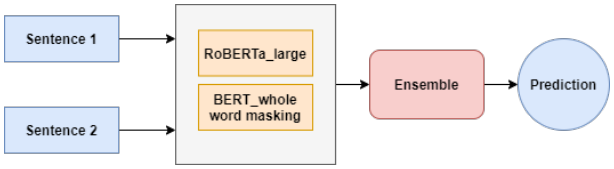

Figure 1: An illustration of the workflow of the proposed ensembling technique.

4 Experimentation and Result

For the sake of this research, accuracy evaluation metric has used to evaluate the performance of the implemented model according to the shared task instruction.

The results of this work obtained during the post-evaluation time. As presented in Table 3, several models were experimented to tackle sub-taskA (Validation) and their hyperparameters that used. The hyper-parameters are used during the training step for the best performs model, we have trained the implemented model for 8 epochs along with 8 Batch size and learning rate of 5e-6. A traditional machine learning Random forest (RF) was trained using count vectors as a baseline model for the results. Table 4 presents the reported results of the evaluation of the performed models. It's obvious according to the results of the performed models have presented in Table 4 that the transformers have different results, for instance, ELMO and USE achieved an accuracy of 58.1%, 67.3% respectively. However, BERT achieved an accuracy of 89.5% and the best performance models are RoBERTa and Bert-WWM with an accuracy of 93.3% and 93.7 respectively including pseudo label technique. As well, it is figured out the best performance for a model was without the use of data pre-processing. Furthermore, the proposed ensembling technique results outperforms the baseline model and gained 93.9% accuracy.

Model	Epochs	Batch size	Learning rate(lr)
ELMO	5	256	2e-5
USE	5	256	2e-5
BERT	3	10	2e-5
XLNET	3	10	2e-5
RoBERTa	8	8	5e-6
BERT-WWM	8	8	5e-6

Table 3: Implemented Models Hyper-parameters

Model	Accuracy
Ensemble technique	**93.9%**
BERT-WWM by Devlin et al. (2018)	93.7%
RoBERTa by Liu et al. (2019)	93.3%
BERT by Devlin et al. (2018)	89.5%
XLNET by (Yang et al., 2019)	72%
USE by (Cer et al., 2018)	67.3%
ELMO by (Peters et al., 2018)	58.1%
RF (baseline) by (Breiman, 2001)	60.43%

Table 4: The performance Results of the Implemented Models

5 Conclusion

This paper presented NLP@JUST teamwork for SemEval-2020 Task 4 competition that related to commonsense validation and explanation (ComVE). The participation is described for the sub-taskA that aims to recognize if the text is common sense or not. The impact of performing transformers (*i.e. Bert, XLNet, and Roberta*) was investigated. It is found that BERT-WWM and Roberta-large obtained the highest accuracy on a test evaluation, also, we adopted those two models to create ensemble technique and produce the final prediction. The evaluation results showed that our implemented technique achieved an accuracy of 93.9%.

References

Leo Breiman. 2001. Random forests. *Mach. Learn.*

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv: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*.

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. *arXiv preprint arXiv:1907.11692*.

Steven L Lytinen. 2005. Artificial intelligence: Natural language processing. *Van Nostrand's Scientific Encyclopedia*.

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*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Shane Storks, Qiaozi Gao, and Joyce Y Chai. 2019. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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. *arXiv preprint arXiv:1803.00191*.

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.

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.

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