daminglu123 at SemEval-2022 Task 2: Using BERT and LSTM to Do Text Classification

Daming Lu
ByteDance
daming.lu@bytedance.com

Abstract

Multiword expressions (MWEs) or idiomaticity are a common phenomenon in natural languages. Current pre-trained language models cannot effectively capture the meaning of these MWEs. The reason is that two single words, after combined together, could have an abruptly different meaning than the compositionality of the meanings of each word, whereas pre-trained language models reply on words' compositionality. We propose an improved method of adding an LSTM layer to the mBERT model to get better results on a text classification task (Subtask A). Our result is slightly better than the baseline. We also tried adding TextCNN to mBERT and adding both LSTM and TextCNN to mBERT. We participate in SubTask A and find that adding only LSTM gives the best performance.

1 Introduction

Machine learning has made deep impacts on various areas, such as computer vision (He et al., 2015, 2017; Lu, 2018), computational biology (Jumper et al., 2021; Huang et al., 2019; Lu, 2010, 2009), and natural language processing (Yang et al., 2019b; Lewis et al., 2019; Madabushi et al., 2020) . In natural language processing, large pre-trained models are prevailing and have achieved great successes. Models such as BERT (Devlin et al., 2018), RoBERTA (Liu et al., 2019), XLNet (Yang et al., 2019a), ALBERT (Lan et al., 2020), Ernie (Sun et al., 2019), etc. performed pretty well in tasks such as sentiment analysis, commonsense reasoning (Lin et al., 2019; Lu, 2020), QA system (Chen and Yih, 2020; Yu et al., 2015) and many other tasks. However, these models are not good at certain tasks such as assessing humor and capturing idiomaticity. This shortcoming is largely due to natural languages' flexibility.

In this paper, we focus on how to use large pre-trained language models to determine whether a multiword expression (MWE) has a trivial meaning (Tayyar Madabushi et al., 2022), a.k.a, the compositionality of each word's meaning, or it is an idiomatic usage. We use the dataset provided in (Tayyar Madabushi et al., 2021). In the training set, the target MWE is given. The previous sentence, the target sentence and the next sentence are also given. We need to decide if the MWE has an idiomatic meaning or its meaning is trivial. This task then can be treated as a text classification problem.

The rest part of this paper is organized as follows:

- We first introduce the dataset and the task with details.
- Then we describe how we built up our pipeline with BERT, LSTM and TextCNN.
- We give our results in section 4.
- Lastly, we provide our discussion in section 5.

2 Dataset and Task

As mentioned in (Tayyar Madabushi et al., 2021), the dataset for Subtask A consists of naturally occurring (target) sentences, previous sentences and next sentences. The target sentence contains potentially idiomatic MWEs annotated with a fine-grained set of meanings: compositional meaning and idiomatic meaning(s). Table 1 shows two samples from the training data. One has an idiomatic expression, and the other not.

3 Methods

Our core pre-trained language model is mBERT (Wolf et al., 2020). We chose mBERT over BERT hoping that it could better fit the task's multi-language specification. In traditional methods, n-gram was used to detect and group the MWEs. In

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 186 - 189
July 14-15, 2022 ©2022 Association for Computational Linguistics

Table 1: Sample data for Subtask A.

previous sentence	target sentence	next sentence	target MWE	label (0 means idiomatic)
"The job has traditionally been nonpolitical, but Mrs. Trump's decision to hire a Trump Organization employee added partisanship to the role, even though Mr. Harleth tried to frame his work there as one stop in a long career in the hospitality industry."	"The White House job was well compensated — former chief ushers say salaries run in the $200,000 range — but the days are long, particularly if the president is an early riser or a night owl; Mr. Trump was both."	Mr. Biden is not a morning person, people familiar with his schedule say.)	night owl	0
Demography expert Piotr Szukalski told Dziennik Gazeta Prawna he thinks that deep concerns about the spread of the coronavirus are to blame.	Minister of Family and Social Policy Marlena Malag ascribed the high death rate to the pandemic and said it would take a long time for the current government program of family benefits intended to boost the birth rate to reverse the negative trend.	"Commenting on data the state agency Statistics Poland released in December for 11 months of 2020, economist Rafal Mundry said the number of deaths was the highest since World War II, and the number of births the lowest in 15 years."	birth rate	1

our methods, we tried to use either LSTM (Hochreiter and Schmidhuber, 1997) or TextCNN (Kim, 2014) to capture the MWEs. We concatenate LSTM or TextCNN to mBERT in order to increase the performance.

3.1 LSTM

Unlike RNN (Jordan, 1997), LSTM is good at remembering only the important parts of a sentence. We hope it can help us group up the MWEs and improve the performance. We add a bidirectional LSTM layer at the output of the sequential transformers. The bidirectional LSTM layer was initialized as 1-layer and bidirectional, with a dropout of 0.1.

3.2 TextCNN

Similar to traditional CNN (Schmidhuber, 2015) in computer vision, TextCNN (Kim, 2014) extracts features from a small area of text. We suppose this layer can help us detect the span of the MWEs so that performance can be improved.

4 Results

We use the mBERT with 12 hidden layers. We did experiments on dropouts with 0.1 and 0.2. As mentioned in Section 3, we explored of adding either a LSTM or a CNN to the final fully connected layer of the transformer from mBERT. Table 2 provides our experiments and results. We were expecting that mBERT + TextCNN could give us the best results. But it turned out that mBERT + LSTM performs best for Subtask A among our experiments. The author has put the code for this paper on GitHub[1].

[1] https://github.com/daming-lu/semeval_2022_task2_sub_a

Table 2: Subtask A Experiment Results

Method	Zero-Shot	One-Shot
mBERT	0.6448	0.6987
+LSTM, dp=0.1	0.6546	0.6998
+LSTM, dp=0.2	0.6333	0.6613
+TextCNN, dp=0.1	0.6501	0.6827
+TextCNN, dp=0.2	0.6254	0.6309
+TextCNN+LSTM	0.6502	0.6977
+LSTM, dp=0.1(**test**)	0.654	0.704

5 Discussion

One reason that our method does not boost the performance a lot might be that we add the LSTM or TextCNN to the end, whose effect is limited to the whole pipeline. Another new method, according to (Gao et al., 2021), is that we can turn this classification problem into a masked word problem. In PROMPT, it claims the integration is more genuine, but choosing the prompt could be technical.

Another important reason is overfitting. We tried to increase dropout from 0.1 to 0.2 in order to get rid of overfitting. But the effect was opposite. According to (Tan et al., 2015), adding LSTM could boost question answering tasks, whereas our task is in fact a text classification. This might be the reason of the tiny improvement.

Acknowledgements

The author would like to thank Dr. Harish Tayyar Madabushi for helpful feedback and suggestions.

References

Danqi Chen and Wen-tau Yih. 2020. Open-domain question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 34–37.

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*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. *CoRR*, abs/1703.06870.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Liang Huang, He Zhang, Dezhong Deng, Kai Zhao, Kaibo Liu, David A Hendrix, and David H Mathews. 2019. LinearFold: linear-time approximate RNA folding by 5'-to-3' dynamic programming and beam search. *Bioinformatics*, 35(14):i295–i304.

Michael I. Jordan. 1997. Serial order: A parallel distributed processing approach. *Advances in psychology*, 121:471–495.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

Yoon Kim. 2014. Convolutional neural networks for sentence classification.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. *arXiv preprint arXiv:1909.02151*.

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*.

Daming Lu. 2009. A combined motif discovery method.

Daming Lu. 2010. A gibbs sampling algorithm for motif discovery using a linear mixed model. In *Proceedings of the International Symposium on Biocomputing*, pages 1–6.

Daming Lu. 2018. Use online dictionary learning to get parts-based decomposition of noisy data. In *17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018, Orlando, FL, USA, December 17-20, 2018*, pages 1492–1494. IEEE.

Daming Lu. 2020. Masked reasoner at SemEval-2020 task 4: Fine-tuning RoBERTa for commonsense reasoning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 411–414, Barcelona (online). International Committee for Computational Linguistics.

Harish Tayyar Madabushi, Elena Kochkina, and Michael Castelle. 2020. Cost-sensitive bert for generalisable sentence classification with imbalanced data. *arXiv preprint arXiv:2003.11563*.

Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2019. Ernie 2.0: A continual pre-training framework for language understanding.

Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2015. Lstm-based deep learning models for non-factoid answer selection. *arXiv preprint arXiv:1511.04108*.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

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

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Zhou Yu, Alexandros Papangelis, and Alexander I. Rudnicky. 2015. Ticktock: A non-goal-oriented multimodal dialog system with engagement awareness. In *AAAI Spring Symposia*.