OCHADAI at SemEval-2022 Task 2: Adversarial Training for Multilingual Idiomaticity Detection

Lis Kanashiro Pereira[1], Ichiro Kobayashi[2]
Ochanomizu University
kanashiro.pereira@ocha.ac.jp[1], koba@is.ocha.ac.jp[2]

Abstract

We propose a multilingual adversarial training model for determining whether a sentence contains an idiomatic expression. Given that a key challenge with this task is the limited size of annotated data, our model relies on pre-trained contextual representations from different multilingual state-of-the-art transformer-based language models (i.e., multilingual BERT and XLM-RoBERTa), and on adversarial training, a training method for further enhancing model generalization and robustness. Without relying on any human-crafted features, knowledge bases, or additional datasets other than the target datasets, our model achieved competitive results and ranked 6th place in SubTask A (zero-shot) setting and 15th place in SubTask A (one-shot) setting.

1 Introduction

Large-scale pre-trained language models such as BERT (Devlin et al., 2019) have achieved great success in a wide range of natural language processing (NLP) tasks. However, more recent studies show that even such contextual models have a limited ability to capture idiomaticity (Garcia et al., 2021). Idiomatic expressions denote a group of words that behave as single words to some extent. Their linguistic behavior cannot be inferred from the characteristics of their components, and still pose a challenge to natural language processing (NLP) systems.

This paper describes the system developed by the OCHADAI team for SemEval-2022 Task 2 - Multilingual Idiomaticity Detection and Sentence Embedding (Tayyar Madabushi et al., 2022). Given that a key challenge in this task is the limited size of annotated data, we follow best practices from recent work on enhancing model generalization and robustness and propose a model ensemble that leverages multilingual pre-trained representations and adversarial training. Our model ranked 6th

Setting	Language	Train	Dev	Eval	Test
zero-shot	English	3,327	-	-	-
	Portuguese	1,164	-	-	-
	Galician	0	-	-	-
one-shot	English	87	466	483	916
	Portuguese	53	273	279	713
	Galician	0	0	0	713

Table 1: Summary of the SemEval 2022 Task 2 Subtask A dataset. Note that the dev, eval and test sets are used in both settings.

on SubTask A (zero-shot), and 15th on SubTask A (one-shot).

2 Task Description

SemEval-2021 Task 2 SubTask A consists of a binary classification task that requires classifying sentences with a target multiword expression (MWE) into either "Idiomatic" or "Literal" across English, Portuguese and Galician (Tayyar Madabushi et al., 2021). Further, it is subdivided into two settings to better test models' ability to generalize: zero-shot and one-shot. In the zero-shot" setting, multiword expressions (potentially idiomatic phrases), in the training set are completely disjoint from those in the test and development sets. In the "one-shot" setting, one positive and one negative training examples are included for each MWE in the test and development sets. Note that the actual examples in the training data are different from those in the test and development sets in both settings. Only the datasets provided by the organizers are allowed to train the models. Participants can use only the data provided for the zero-shot setting to train the zero-shot model. However, participants were allowed to use data provided for both settings to train models in the one-shot setting. The statistics of the corpus are presented in Table 1. Our team submitted results for both settings, and the next section outlines the overview of our model.

Setting	Language	Sentence	Target MWE	Label
zero-shot	English	This song is about unconditionally supporting someone you love. This is a **love song**. Let's be there for each other.	love song	1
one-shot	Portuguese	Estamos honrando o teto e construindo as paredes", afirmou. Em outro momento, voltando ao tema do impasse do Orçamento, ele reforçou que a busca é por uma solução, que a briga tem "mérito" e que "sempre, em grande rebanho, tem uma **ovelha negra**"."Mas não é o Congresso nem o grosso do ministério", completou.	ovelha negra	0
one-shot	Galician	Non podemos abandonalos á súa sorte, porque sen mozos e mozas no campo e sen a súa actividade agraria suporía, entre outras cousas, a deslocalización da produción e a dependencia alimentaria Máis alá das políticas comunitarias, ¿cres que poden desenvolverse alternativas para a **xente nova** a partir de medidas municipais, autonómicas ou estatais? Estase a facer?	xente nova	0

Table 2: Example sentences and labels for Subtask A. Note that "Idiomatic" is assigned the label 0 in the dataset and "non-idiomatic" (including proper nouns) are assigned the label 1.

3 System Overview

We focus on exploring different training techniques using BERT and RoBERTa, given their superior performance on a wide range of NLP tasks. Each text encoder and training method used in our model are detailed below.

3.1 Text Encoders

M-BERT (Devlin et al., 2019): We use the M-BERT$_{\text{BASE}}$ model released by the authors. It is pre-trained on the top 104 languages with the largest Wikipedia using a masked language modeling (MLM) objective. This model is case sensitive: it makes a difference between english and English.
XLM-R (Conneau et al., 2019): XLM-RoBERTa (XLM-R) is a multilingual version of RoBERTa. It is pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages. XLM-R has been shown to perform particularly well on low-resource languages, such as Swahili and Urdu. We use the XLM-R$_{\text{LARGE}}$ model released by the authors.

3.2 Training Procedures

Standard fine-tuning: This is the standard fine-tuning procedure where we fine-tune BERT and RoBERTa on each training setting-specific data.
Adversarial training (ADV): Adversarial training has proven effective in improving model generalization and robustness in computer vision (Madry et al., 2017; Goodfellow et al., 2014) and more recently in NLP (Zhu et al., 2019; Jiang et al., 2019; Cheng et al., 2019; Liu et al., 2020a; Pereira et al., 2020). It works by augmenting the input with a small perturbation that maximizes the adversarial loss:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D}[\max_{\delta} l(f(x+\delta;\theta),y)] \quad (1)$$

where the inner maximization can be solved by projected gradient descent (Madry et al., 2017). Recently, adversarial training has been successfully applied to NLP as well (Zhu et al., 2019; Jiang et al., 2019; Pereira et al., 2020). In our experiments, we use SMART (Jiang et al., 2019), which instead regularizes the standard training objective using *virtual adversarial training* (Miyato et al., 2018):

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D}[l(f(x;\theta),y)+ \\ \alpha \max_{\delta} l(f(x+\delta;\theta), f(x;\theta))] \quad (2)$$

Effectively, the adversarial term encourages smoothness in the input neighborhood, and α is a hyperparameter that controls the trade-off between standard errors and adversarial errors.

4 Experiments

4.1 Implementation Details

Our model implementation is based on the MT-DNN framework (Liu et al., 2019a, 2020b). We use BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2019) as the text encoders. We used ADAM (Kingma and Ba, 2015) as our optimizer with a learning rate in the range $\in \{8 \times 10^{-6}, 9 \times 10^{-6}, 1 \times 10^{-5}\}$ and a batch size $\in \{8, 16, 32\}$. The maximum number of epochs was set to 10. A linear learning rate decay schedule with warm-up over 0.1 was used, unless stated otherwise. To avoid gradient exploding, we clipped the gradient norm within 1. All the texts were tokenized using wordpieces and were chopped to spans no longer than 512 tokens. During adversarial training, we follow (Jiang et al., 2019) and set the perturbation size to 1×10^{-5}, the step size to 1×10^{-3}, and to 1×10^{-5} the variance for initializing the perturbation. The number of projected gradient steps and the α parameter (Equation 2) were both set to 1.

We follow (Devlin et al., 2019) and (Liu et al., 2019b), and set the first token as the [CLS] token and the <s> token, respectively, when encoding the input on BERT and RoBERTa, respectively. We separate the input sentence and the target expression with the special token [SEP] and </s> for BERT and RoBERTa, respectively. e.g. [CLS] Ben Salmon is a committed *night owl* with an undying devotion to discovering new music.,"He lives in the great state of Oregon, where he hosts a killerradio show and obsesses about Kentucky basketball from afar. [SEP] *night owl* [SEP].

For both settings (zero-shot and one-shot), we used the dev dataset released by organizers to fine-tune the model's hyperparameters.

4.2 Main Results

Submitted systems were evaluated in terms of F1-score. The systems were ranked from highest F1-score score to lowest. We built several models that use different text encoders and different training methods, as described in Section 3. See Table 3 for the results.

First, we observe that models that use adversarial training obtained better performance overall, without using any additional knowledge source, and without using any additional dataset other than the target task datasets. These results suggest that adversarial training leads to a more robust model and helps generalize better on unseen data. For the zero-shot setting, the model that uses XLM-R as the text encoder and adversarial training performed better than M-BERT on the development set. Thus, we decided to submit this model's results on the test set. It obtained a test set F1-score of 0.7457, and ranked 6th among all participating systems. On the other hand, on the one-shot setting, M-BERT performed better than XLM-R on the development set. Again, M-BERT with adversarial training performed better than vanilla fine-tuning. This model obtained an F1-score of 0.6573 on the test set, and ranked 15th among all participating systems.

5 Conclusion

We proposed a simple and efficient model for multilingual idiomaticity detection. Our experiments demonstrated that it achieves competitive results on both zero-shot and one-shot settings, without relying on any additional resource other than the target task dataset. Although in this paper we focused on the multilingual idiomaticity detection task, our model can be generalized to solve other

Method	zero-shot F1	one-shot F1
Dev		
Standard$_{\text{XLM-R}_{\text{LARGE}}}$	0.7076	0.7769
SMART$_{\text{XLM-R}_{\text{LARGE}}}$	0.7378	0.7958
Standard$_{\text{M-BERT}_{\text{BASE}}}$	0.5540	0.8462
SMART$_{\text{M-BERT}_{\text{BASE}}}$	0.6200	0.8568
Test		
SMART$_{\text{XLM-R}_{\text{LARGE}}}$	0.7457	-
SMART$_{\text{M-BERT}_{\text{BASE}}}$	-	0.6573

Table 3: Comparison of standard and adversarial training in zero-shot evaluation on various natural language inference datasets, where the standard BERT$_{\text{BASE}}$ model is fine-tuned on the MNLI training data.

downstream tasks as well, and we will explore this direction as future work.

Acknowledgements

References

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Probing for idiomaticity in vector space models. In *Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL).

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2019. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. *arXiv preprint arXiv:1911.03437*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR (Poster) 2015*.

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020a. Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496.

Xiaodong Liu, Yu Wang, Jianshu Ji, Hao Cheng, Xueyun Zhu, Emmanuel Awa, Pengcheng He, Weizhu Chen, Hoifung Poon, Guihong Cao, and Jianfeng Gao. 2020b. The microsoft toolkit of multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:2002.07972*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993.

Lis Pereira, Xiaodong Liu, Fei Cheng, Masayuki Asahara, and Ichiro Kobayashi. 2020. Adversarial training for commonsense inference. *arXiv preprint arXiv:2005.08156*.

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*, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Thomas Goldstein, and Jingjing Liu. 2019. Freelb: Enhanced adversarial training for language understanding. *arXiv preprint arXiv:1909.11764*.