ZhichunRoad at SemEval-2022 Task 2: Adversarial Training and Contrastive Learning for Multiword Representations

Xuange Cui , Wei Xiong , Songlin Wang
JD.com, Beijing, China
{cuixuange,xiongwei9,wangsonglin3}@jd.com

Abstract

This paper presents our contribution to the SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding. We explore the impact of three different pre-trained multilingual language models in the SubTaskA. By enhancing the model generalization and robustness, we use the exponential moving average (EMA) method and the adversarial attack strategy. In SubTaskB, we add an effective cross-attention module for modeling the relationships of two sentences. We jointly train the model with a contrastive learning objective and employ a momentum contrast to enlarge the number of negative pairs. Additionally, we use the alignment and uniformity properties to measure the quality of sentence embeddings. Our approach obtained competitive results in both subtasks.

1 Introduction

In recent years, the pre-trained models have been widely used and play a vital role in the natural language processing tasks. The success of language models relies on huge amounts of unlabeled data and the useful representation layers that are designed to draw on information from the surrounding context (Devlin et al., 2018;Nedumpozhimana and Kelleher, 2021). However, more recent studies show that even state-of-the-art pre-trained contextual models (e.g. BERT) can't accurately represent idiomatic expressions (Yu and Ettinger, 2020;Garcia et al., 2021). One reason for this is that many expressions can be used both literally and idiomatically.

Specifically, the size of vocabulary can't increase indefinitely, which makes representing idiomatic phrases particularly challenging (Shwartz, 2021;Tayyar Madabushi et al., 2021). Idioms occur in almost all languages, to distinguish whether an expression has an idiomatic sense would leverage both cross-lingual models and multiword expressions (MWEs). The SemEval 2022 Task 2(Tayyar Madabushi et al., 2022) is aimed at detecting and representing MWEs and presents a novel multilingual dataset across English, Portuguese and Galician. And this task consists of two different subtasks to evaluate the model's ability to identify and capture idiomaticity.

SubTask	Train	Dev	Test	Desc
zero-shot	4492	740	763	binary classification
one-shot	4492	740	763	
pre-train	24498	2000	3827	semantic similarity
fine-tune	6573	2182	2263	

Table 1: The statistics of datasets.

Our contributions can be summarized as follows: 1) We choose three transformer-based language models from the XTREME LeaderBoard (Hu et al., 2020), and compare the effectiveness of mBERT$_{base}$ (Devlin et al., 2018), XLM-R$_{base}$(Conneau et al., 2020), and InfoXLM$_{base}$(Chi et al., 2021). 2) We adopt the exponential moving average method (EMA) and adversarial training strategy to improve the model's generalization and robustness. We achieved considerably performance gain of 2.91%, 3.87% over the baseline solution, and ranked 12th in the zero-shot settings and 4th in the one-shot settings. 3) With finetuning on the supervised target datasets, we use cross-attention module and jointly train the model with an extra contrastive loss layer on top of the BERT encoder. Our approach achieved an 8.22%, 4.5% improvement compared to the baseline solution, and ranked top-4 in the pre-train and fine-tune settings. We release the source code and pre-trained models associated with this work. [1]

Moreover, we find that adversarial training can

[1] https://github.com/cuixuage/SemEval2022-Task2

achieve good performance by setting the appropriate batch size. In the SubTaskB, we show that joint training regularizes the sentence embeddings' anisotropic space to be more uniform but also suffers a degeneration in alignment slightly. The trade-off between the alignment and uniformity (Wang and Isola, 2020) indicates that perfect alignment and perfect uniformity are likely hard to simultaneously achieve in practice.

2 Background

SemEval-2022 Task 2(Tayyar Madabushi et al., 2022) provides two subtasks. Subtask A consists of a binary classification task aimed at determining whether a sentence contains an idiomatic expression. The sample of the dataset consists of the previous sentence, target sentence, next sentence, and MWE. The target sentence contains the potentially idiomatic MWE, and the label of 0 indicates "Idiomatic" and the label of 1 indicates "non-idiomatic". Our model receives the context sentences as input in the zero-shot setting, and receives the target sentence by adding the MWE as an additional feature in the one-shot setting. This is based on the results presented in the dataset paper (Tayyar Madabushi et al., 2021).

SubtaskB consists of a novel task which requires the model to output the correct Semantic Text Similarity (STS) scores. The task is designed to test a model's ability to generate sentence embeddings that accurately represent sentences regardless of whether or not they contain idiomatic expressions. When evaluating the trained model, we first obtain the sentence embeddings, then we calculate the Spearman correlation between the cosine similarity scores of sentence embeddings and the gold labels. The statistics of the corpus are shown in Table 1. Our team participated in both subtasks, and the next section will introduce an overview of our system.

3 System Overview

We focus on comparing the impact of different training techniques adopted in our system. In this section, we first present the BERT-like text encoder, then we introduce several strategies for improving models' robustness. Finally, we talk about the design of the cross-attention module and the jointly training way of incorporating supervised signals and unsupervised signals.

3.1 Transformer-based Models

In the zero-shot and one-shot settings, we compare several pre-trained multilingual language models from the XTREME Leaderboard[2] as the text encoder . The models shown below are also available on the hugging-face website[3].

mBert$_{base}$,the bert-base-multilingual-cased model is pre-trained on the top 104 languages with the Wikipedia dataset, and consists of 12-layer, 768-hidden, 12-heads, 109M parameters and a shared vocabulary size of 110000 (Devlin et al., 2018).

XLM-R$_{base}$,the xlm-roberta-base model consists of 100 languages and pre-trained with filtered CommonCrawl dataset, and consists of 12-layer, 768-hidden, 12-heads, $\tilde{2}$70M parameters and a shared vocabulary size of 250002 (Conneau et al., 2020).

InfoXLM$_{base}$,we use the "microsoft/infoxlm-base" model containing 94 languages and pre-trained with CCNet dataset, and has the same configurations of XLM-R and a shared vocabulary size of 250002 (Chi et al., 2021).

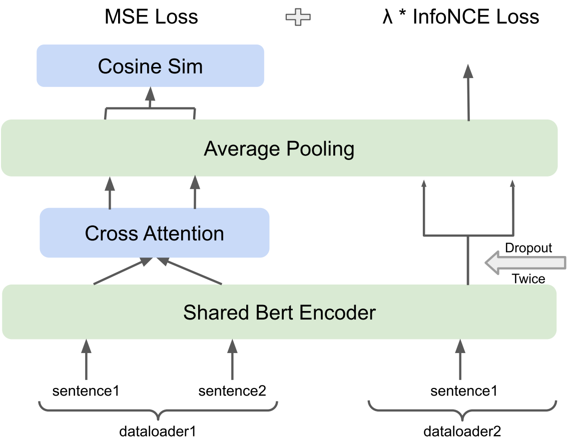

Figure 1: Incorporating supervised and unsupervised signals. MSE Loss: the mean squared error, InfoNCE Loss: the contrastive objective.

3.2 Training Procedures

There are two ways of enhancing the model generalization and robustness.

Exponential Moving Average Our model uses EMA to smooth the trained parameters. Evaluations that use averaged parameters sometimes produce significantly better results than the final trained values. Formally, we define the smoothed

[2] https://sites.research.google/xtreme
[3] https://huggingface.co/models

SubTaskB	Model	ALL Data	Idiom Data	STS Data
pre-train	mBert$_{base}$	53.90	21.87	80.82
	mBert$^{\diamondsuit}_{base}$	55.58	27.18	82.09
	mBert$^{\clubsuit}_{base}$	56.32	28.26	83.59
fine-tune	mBert$_{base}$	62.29	34.59	52.29
	mBert$^{\diamondsuit}_{base}$	63.16	36.95	53.49
	mBert$^{\clubsuit}_{base}$	64.01	39.56	56.15

Table 2: Performance of Our Approach on the Sentence Representation Task. We report the Spearman correlation × 100 on the test sets, $\diamondsuit$: jointly train the model with the contrastive objective, $\clubsuit$: jointly train the model with the cross-attention module and the contrastive objective.

variables and trained variables as θ_s and θ_t, EMA decay weight as: η. After each training step, we update θ_s by:

$$\theta_s \leftarrow \eta\theta_s + (1-\eta)\theta_t \tag{1}$$

Adversarial Training Recently, adversarial attack has been widely applied in computer vision and natural language processing (Yan et al., 2021). Many works use it during fine-tuning, because computing adversarial perturbations relies on supervised signals. We explore the influence of adversarial training strategies with different batch size, and compare the FGSM (Goodfellow et al., 2015), PGD (Madry et al., 2019), FREELB (Zhu et al., 2020) and SMART (Jiang et al., 2020) methods in the zero-shot and one-shot settings. It works by augmenting the input with a small perturbation that maximizes the adversarial loss:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\Delta x\in\Omega} L(x+\Delta x, y;\theta)\right] \tag{2}$$

where the $\mathcal{D}$ is dataset, x is input, y is the gold label, θ is the model parameters, $L(x, y; \theta)$ is the loss function and Δx is the perturbation. In our experiments, we adopt SMART method in zero-shot setting, and FREELB method in one-shot setting. These choices are based on actual performance.

3.3 Sentence Representation

Reimers and Gurevych (2019) propose a siamese architecture with a shared BERT encoder to compute the sentence representations for each input text. By making use of unlabeled texts, SimCSE (Gao et al., 2021) proposes an unsupervised contrastive learning method to alleviate the collapse issue of BERT. Compared to unsupervised SimCSE, we use extra supervised signals during training. Our approach is mainly inspired by ConSERT (Yan et al., 2021) and EsimCSE (Wu et al., 2021). As shown in Figure 1, there are two major objectives and an extra cross-attention module to exchange information with the token-wise embeddings.

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{mse}} + \lambda\mathcal{L}_{\text{con}} \tag{3}$$

the λ is a hyperparameter to balance two objectives. $\mathcal{L}_{\text{mse}}$ is Mean Squared Error, $\mathcal{L}_{\text{con}}$ is Contrastive Loss.

During training, each data point is trained to find out its counterpart among $(N-1)$ from in-batch negative samples and the queue of data samples. The samples in the queue are progressively replaced (He et al., 2020).

$$-\log \frac{e^{\mathrm{sim}\left(\mathbf{h}_i,\mathbf{h}_i^{+}\right)/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}\left(\mathbf{h}_i,\mathbf{h}_j^{+}\right)/\tau} + \sum_{q=1}^{Q} e^{\mathrm{sim}\left(\mathbf{h}_i,\mathbf{h}_q^{+}\right)/\tau}} \tag{4}$$

The h_* is the sentence representation, where h_i and h_i^+ are semantically related. The h_q^+ is denotes a sentence embedding in the momentum-updated queue. And the Q is the size of the queue, $sim(h1, h2)$ is the cosine similarity scores of sentence representations, τ is a temperature hyperparameter. In the end, we average the all N Li losses to calculate the contrastive loss $\mathcal{C}_{\text{con}}$.

4 Experiments

4.1 Settings

We use InfoXLM$_{base}$ (Chi et al., 2021) as the text encoder, the EMA decay weight is set to 0.999, the learning rate is set to 2e-5 with warmup ratio over 10% in the SubTaskA. We compare the impact of batch size $\in$ 16, 32, 64 with different adversarial training strategies. By default, We set ε to 1.0 in FGM, set K steps to 3 in PGD, FREELB and SMART. That means calculate 3 steps in the adversarial attack. We adopt λ as 0.5 and μ as 0.2 to smooth the logits and embeddings in the SMART method (Jiang et al., 2020). We use SMART method and batches of size 16 in zero-shot setting, FREELB method and batches of size 32 in one-shot setting.

SubTask	Model	Practice	Post-Eval
zero-shot	mBert$_{base}$	68.71	-
	infoxlm$^{\diamondsuit}_{base}$	76.21	68.31
one-shot	mBert$_{base}$	84.77	-
	infoxlm$^{\clubsuit}_{base}$	94.07	90.33

Table 3: Performance of Our Approach on the Idiomaticity Detection Task. We report the F1 Score × 100 on the dev and test sets, ♢: set bath size to 16 and use SMART, ♣: set bath size to 32 and use FREELB.

Method	Practice	Post-Eval
mBert$_{base}$	68.71	-
InfoXLM$_{base}$	73.10	65.2
+EMA	75.75	67.85
+EMA+SMART	76.21	68.31

Table 4: The effect of different strategies and keep accumulating from top to bottom. We report the dev-F1 Score × 100 in zero-shot setting.

In the SubTaskB, we use mBERT$_{base}$ (Devlin et al., 2018) as the text encoder, set batch size to 32 and set warmup ratio to 10%. During the jointly training, λ is set to 0.15 and τ is set to 0.05 that used in the $\mathcal{L}_{\text{con}}$. We use the dev set of STS-B and ASSIN2 to tune the hyperparameter and evaluate the model every 250 steps during training. The best checkpoint is saved for testing, we further discuss the results of our experiments in the subsequent section.

4.2 Main Results

Our submitted results were evaluated on F1 Score in SubTaskA, and Spearman correlation in SubTaskB. We jointly train the model with contrastive objective and the supervised signals on the Semantic Text Similarity dataset, including STSBenchmark and ASSIN2 datasets. We compare several models as the text encoder and different training methods, as described in Section 3. The main results shown in Table 2 and Table 3. As shown in Table 3, we achieve a performance gain of 2.91%, 3.87% over the baseline solution by finetuning the InfoXLM$_{base}$ model with using EMA method and adversarial training. In the Table 2, our approach achieves a 8.22%, 4.5% improvement compared to the baseline solution that indicates the usefulness of the cross-attention module and jointly training way. In the next section, we study the effect of different strategies.

5 Ablation Studies

5.1 Effect of Pre-trained Models

We investigate the impact of adopting different multi-lingual models in the zero-shot setting. In Figure 2, we show the results of different language models fine-tuning in 50 epochs. We find that the best f1 score on validation dataset is provided by InfoXLM$_{base}$ (Chi et al., 2021).

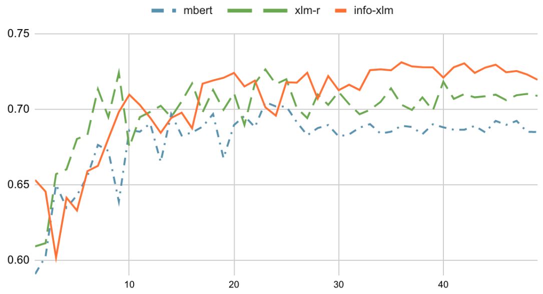

Figure 2: The fine-tuning of multi-lingual language models. We report the dev-F1 Score in zero-shot setting.

5.2 Effect of Training Techniques

As shown in Figure 3, we set bath size to 16 and use 0.999 as the EMA decay weight to obtain the best score. In the zero-shot and one-shot settings, we find that the performance is extremely sensitive to the batch size. And with the benefit of smoothing performance, using the EMA method can improve the model robustness when evaluating the trained model.

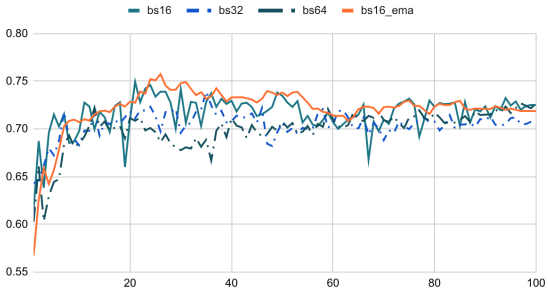

Figure 3: The batch size with EMA method. We report the dev-F1 Score in zero-shot setting.

The experimental results of adversarial training are presented in Figure 4. We set the size of mini-batch to 16 and use SMART in the zero-shot setting, and the batch size is set to 32 and use FREELB as adversarial attack in one-shot setting. We observe that the SMART and FREELB strategies have better performance than FGM and PGD strategies.

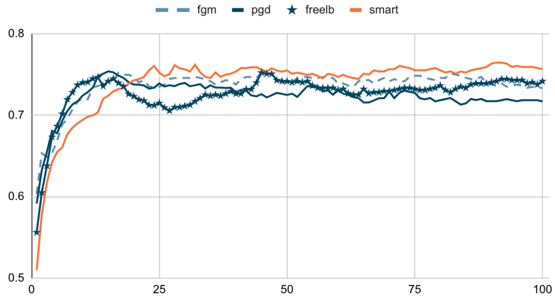

Figure 4: The performance of different adversarial attack strategies. We report the dev-F1 Score in zero-shot setting.

As presented in Table 5, we explore the impact of InfoXLM$_{base}$ model, smaller batch size, EMA method and adversarial training. These strategies can effectively improve the performance of our approach.

5.3 Effect of Contrastive Learning

In this section, we investigate the contrastive learning how to further improve the performance of sentence representations. As shown in Table 5, we use contrastive learning as unsupervised signals which yields a substantial improvement on the STS-Test dataset. We also use an extra cross-attention layer to achieve a 0.8%, 1.2% improvement in the pre-train and fine-tune settings. The cross attention idea is inspired by Reimers and Gurevych (2019), the paper shows that the Cross-Encoder achieves better performances than Bi-Encoders.

Method	Practice	Post-Eval
mBert$_{base}$	70.33	-
+CrossAttention	70.96	55.94
+ + InfoNCE	71.11	56.09
+ + + MoCo	71.34	56.32

Table 5: The effect of different strategies on the STSTest dataset. We report the Spearman correlation × 100 in pre-train setting.

In general, models which have both better alignment and uniformity obtain better sentence representations, confirming the findings in Wang and Isola (2020). We also evaluate these metrics to measure the quality of learned embeddings, including alignment of the positive pairs and uniformity of the whole representation space. We calculate uniformity on the STS-B and ASSIN2 datasets, and alignment from the positive pairs that have the gold label more than or equal to the number 4.

As shown in Figure 5, we also find that: 1) Though pre-trained embeddings have good alignment, their uniformity is poor, e.g. mBert$_{base}$. 2) Unsupervised SimCSE$_{base}$ (Gao et al., 2021) has better uniformity of pre-trained embeddings than mBert$_{base}$. 3) Jointly training regularizes the sentence embeddings' anisotropic space to be more uniform than others, but also suffers a degeneration in alignment slightly. 4) The trade off between the alignment and uniformity indicates that perfect alignment and perfect uniformity are likely hard to simultaneously achieve in practice.

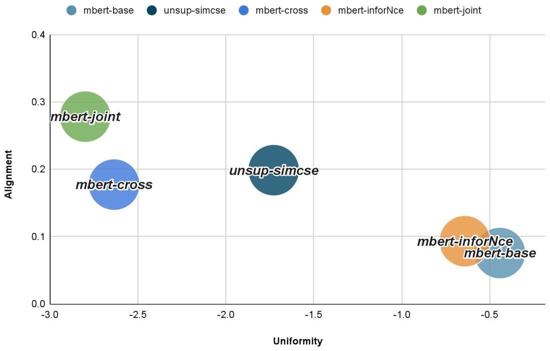

Figure 5: The alignment and uniformity of different pre-trained models. The closer to the origin of the coordinate axis, the better sentence representations.

6 Conclusion and Future Work

In this work, we provide an overview of the combined approach to detect and represent multiword expressions. We use InfoXLM$_{base}$ model as the text encoder and enhance the model generalization and robustness with exponential moving average (EMA) method and the adversarial attack strategy in the SubTaskA. In the SubTaskB, experimental results show that the cross-attention module and the contrastive learning task can considerably improve the performance. Finally, we analyze the alignment and uniformity properties to measure the quality of sentence embeddings. Future work of our system includes: 1) Using the larger pre-trained language models, such as mBert$_{large}$, InfoXLM$_{large}$. 2) Adopting other data augmentation, including Token-Shuffle, Token-Cutoff and Mix-Up.

References

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

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: Main Volume*, pages 3551–3564, Online. Association for Computational Linguistics.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: Robust and efficient fine-tuning for pretrained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190, Online. Association for Computational Linguistics.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2019. Towards deep learning models resistant to adversarial attacks.

Vasudevan Nedumpozhimana and John Kelleher. 2021. Finding BERT's idiomatic key. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 57–62, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Vered Shwartz. 2021. A long hard look at MWEs in the age of language models. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, page 1, Online. Association for Computational Linguistics.

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.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere.

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2021. Esimcse: Enhanced sample building method for contrastive learning of unsupervised sentence embedding.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, Online. Association for Computational Linguistics.

Lang Yu and Allyson Ettinger. 2020. Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4907, Online. Association for Computational Linguistics.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*.