

TADA: Task-Agnostic Dialect Adapters for English

William Held[†] Caleb Ziems[‡] Diyi Yang[‡]
 Georgia Institute of Technology,  Stanford University
 wheld3@gatech.edu

Abstract

Large Language Models, the dominant starting point for Natural Language Processing (NLP) applications, fail at a higher rate for speakers of English dialects other than Standard American English (SAE). Prior work addresses this using task-specific data or synthetic data augmentation, both of which require intervention for each dialect and task pair. This poses a scalability issue that prevents the broad adoption of robust dialectal English NLP. We introduce a simple yet effective method for task-agnostic dialect adaptation by aligning non-SAE dialects using adapters and composing them with task-specific adapters from SAE. Task-Agnostic Dialect Adapters (**TADA**) improve dialectal robustness on 4 dialectal variants of the GLUE benchmark without task-specific supervision.¹

1 Introduction

Large Pretrained Language Models (LLMs; Devlin et al., 2019; Liu et al., 2019; Raffel et al., 2020) have been shown to perform much worse for English dialects other than Standard American English (SAE) (Ziems et al., 2022, 2023). Existing work on dialectal English NLP is task-specific, using manually annotated dialect data (Blodgett et al., 2018; Blevins et al., 2016), weak-supervision (Jørgensen et al., 2016; Jurgens et al., 2017), or data augmentation (Ziems et al., 2022, 2023).

As LLMs become a general-purpose technology, they are applied in an increasing number of scenarios by users who are not formally trained in Machine Learning (Bommasani et al., 2021). Non-experts rarely look beyond accuracy (Yang et al., 2018), making them less likely to value robustness above the cost of training (Ethayarajh and Jurafsky, 2020). Unmitigated dialect bias in this long tail of tasks has the potential to exacerbate harms due to unfair allocation of resources (Bender et al., 2021).

¹We release code for training both traditional and task-agnostic adapters for English dialects on GitHub and finetuned models, adapters, and TADA modules on HuggingFace.

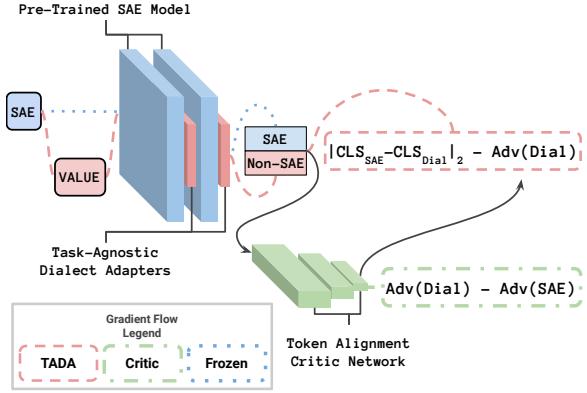


Figure 1: TADA trains adapters with both sequence and token level alignment loss between SAE and a target dialect. When stacked before task-specific SAE adapters, TADA provides dialect robustness for the target task.

Dialectal discrepancies originate in biases in the filtering of LLM pretraining data before finetuning (Gururangan et al., 2022). Despite dialects being definitionally similar, training which enables task-agnostic zero-shot transfer is underexplored relative to potential utility (Bird, 2022). Such task-agnostic transfer methods are natural, practical, and offer a scalable solution for English dialects across the growing spectrum of NLP applications.

This work contributes the first pursuit of these goals with **Task-Agnostic Dialect Adapters (TADA)**. Adapters, bottlenecks placed between transformer layers, provide a parameter-efficient (Houlsby et al., 2019) and composable (Pfeiffer et al., 2020) foundation for task-agnostic dialect adaptation, given the low-resourced nature of most dialects. As shown in Figure 1, TADA modules are trained to align non-SAE dialect inputs with SAE inputs at multiple levels with both a sequence-level contrastive loss and a novel morphosyntactic loss.

We show the empirical effectiveness of TADA on 4 dialect variants of GLUE (Wang et al., 2018) with perturbations from Ziems et al. (2023). We re-

lease TADA as a plug-and-play tool for mitigating dialect discrepancies, launching a scalable pathway to dialect-inclusive English NLP.

2 Related Work

NLP For English Dialects Existing work on NLP for English dialects has largely focused on data collection and weak supervision. Jørgensen et al. (2016) uses online lexicons to provide weak supervision for AAE. Blevins et al. (2016) manually annotates a small dataset and uses domain adaptation methods to enable transfer. Jurgens et al. (2017) collects a geographically diverse set of English data and uses distant supervision signals to annotate a large and representative language ID corpus. Multi-VALUE (Ziems et al., 2022, 2023) develops a data augmentation framework for task-specific training in many common English dialects. Our work proposes a complementary task-agnostic intervention for English NLP.

Cross-Lingual Alignment Cross-lingual alignment has become a common approach for task-agnostic zero-shot transfer across languages. Explicit lexical alignment can be used to learn cross-lingual word embeddings for downstream tasks (Duong et al., 2016; Adams et al., 2017; Artetxe et al., 2018; Grave et al., 2019). More recent work shows that end-to-end models can implicitly learn to align representations (Zoph et al., 2016; Conneau and Lample, 2019; Conneau et al., 2020; Xue et al., 2021). These alignment methods often perform better on highly similar languages, making them theoretically well-suited for dialects. By using explicit alignment with composable modules, our work is the first to explore such techniques for English dialectal NLP.

Adapters A growing body of research has been devoted to finding scalable methods for adapting increasingly large-scale pre-trained models. Houlsby et al. (2019) adapt large models using bottleneck layers (with skip-connection) between each layer. This idea has been extended in many domains (Stickland and Murray, 2019; Pfeiffer et al., 2021; Rebuffi et al., 2017; Lin et al., 2020). Most relevant, Pfeiffer et al. (2020) showed that discrete language modeling adapters and task adapters can be composed for effective cross-lingual multi-task transfer. Our experiments exploit specialized dialectal data augmentation to extend this approach to English dialects using explicit alignment loss.

3 TADA: Task-Agnostic Dialect Adapters

As an initial effort, TADA aims to provide task-agnostic dialect robustness for English NLP. To do so, we build on work from both multilingual NLP and computer vision and apply explicit alignment losses for transfer learning. Concretely, we first generate a synthetic sentence-parallel corpus using the morphosyntactic transformations created by Ziems et al. (2023). Using these parallel sentences, we train TADA to align using a contrastive loss at the sequence level and an adversarial loss at the token level. At test time, TADA modules are stacked with task-specific adapters trained on SAE to improve the dialect performance on the target task without further training.

3.1 Synthetic Parallel Data

While cross-lingual transfer has leveraged the wealth of sentence parallel bi-texts from machine translation to learn alignment, there are no large-scale parallel English dialectal datasets. Therefore, we leverage Multi-VALUE, a rule-based morphosyntactic SAE to a non-SAE translation system to create parallel data (Ziems et al., 2023).

We start with SAE sentences sampled from the Word-in-Context (WiC) Dataset (Pilehvar and Camacho-Collados, 2019). WiC is designed to contain lexically diverse sentences and is sourced from high-quality lexicographer written examples (Miller, 1994; Schuler, 2005). This avoids our alignment modules overfitting to specific vocabulary or noise from low-quality examples. We generate 1,000 such pairs, an amount which could be feasibly replaced with human-translated data.

This data limitation is intentional, as Multi-VALUE could alternatively used to do large-scale pretraining on transformed data (Qian et al., 2022). With smaller data limitations, the data used to train TADA can be manually curated native speakers and linguists to most accurately describe the dialect via minimal pairs (Demszky et al., 2021). Additionally, it opens the potential for TADA to be used for non-English dialects, related languages, and codeswitched variants where small amounts of manually translated data already exists (Diab et al., 2010; Salloum and Habash, 2013; Klubička et al., 2016; Costa-jussà, 2017; Costa-jussà et al., 2018; Popović et al., 2020; Chen et al., 2022; Agarwal et al., 2022; Hamed et al., 2022) Furthermore, using a small amount of data, in combination with a parameter-efficient method, reduces compute costs

Dialect Adaptation Details				AAE Glue Performance							
Approach	Method	Task-Agnostic	Dialect Params.	COLA	MNLI	QNLI	RTE	QQP	SST2	STS-B	Mean
N/A	Finetuning	✓	0	13.5	82.0	89.3	71.8	87.1	92.0	89.9	75.1
N/A	Adapters	✓	0	14.1	83.7	90.3	67.1	86.8	92.1	88.7	74.7
VALUE	Finetuning	✗	$T \times 110M$	19.8	84.9	90.8	74.4	89.6	92.4	90.9	77.5
VALUE	Adapters	✗	$T \times 895K$	40.2	85.8	92.2	73.6	89.7	93.6	90.3	80.8
TADA	Adapters	✓	$895K$	29.5+	84.8+	91.7+	67.2+	88.1+	91.9	89.6+	77.5+

Table 1: **Dialect Adaptation GLUE** results of RoBERTa Base (Liu et al., 2019) for the 7 GLUE Tasks (Matthew’s Corr. for CoLA; Pearson-Spearman Corr. for STS-B; Accuracy for all others). T is the number of target tasks for dialect adaptation. Tasks where TADA improves the performance of task-specific SAE adapters, are marked with +.

as a barrier for dialect speakers to develop and own language technology within their communities (Ahia et al., 2021).

3.2 Contrastive Sequence Alignment

Multilingual NLP has shown that L_2 alignment on small amounts of data can provide competitive performance gains to augmentation using translated data during finetuning (Conneau et al., 2018). This operates on the intuition that similar input representations are likely to lead to similar outputs.

TADA extends this approach to dialects by minimizing the L_2 distance between a frozen representation of an SAE input CLS_{sae} and the TADA representation of a non-SAE input CLS_{dial} :

$$L_{seq} = |\text{CLS}_{sae} - \text{CLS}_{dial}|_2 \quad (1)$$

3.3 Adversarial Morphosyntactic Alignment

Since our translated data is aligned at the sequence level, the contrastive loss is only applied to the **CLS** representations. However the variation, and therefore our ideal alignment procedure, operates at the morphosyntactic level.

Lacking token-level aligned data, we instead pursue morphosyntactic alignment using unsupervised adversarial alignment methods (Zhang et al., 2017; Lample et al., 2018). Since our goal is to capture morphosyntactic differences, we use an adversary which pools the entire sequence using a single-layer transformer (Vaswani et al., 2017) with a two-layer MLP scoring head. A transformer adversary has the expressive capacity to identify misalignment in both individual tokens and their relationships.

We leave the source dialect frozen which has been shown in computer vision to lead to representations that are composable with downstream modules (Tzeng et al., 2017). Given the adversarial scoring network Adv , a frozen SAE representation SAE , and a Non-SAE representation after TADA Dial , we train Adv to maximize:

$$L_{adv} = \text{Adv}(\text{Dial}) - \text{Adv}(\text{SAE}) \quad (2)$$

Then, define the morphosyntactic loss for TADA by minimizing the critic loss from Adv :

$$L_{ms} = -\text{Adv}(\text{Dial}) \quad (3)$$

3.4 Plug-And-Play Application

Finally, we propose a procedure for applying TADA to downstream tasks. We use composable invertible adapters (Pfeiffer et al., 2020) as our starting point. Using the 1,000 sentences from WiC, we train these adapters to minimize the combined contrastive and adversarial loss functions:

$$L_{TADA} = L_{seq} + L_{ms} \quad (4)$$

At test time TADA modules can be stacked behind traditional task adapters (Houlsby et al., 2019). TADA serves to directly align the representations of Non-SAE inputs to the SAE embedding space that these task adapters were trained on. Our experiments show that this consistently improves adapter performance without further training.

4 Evaluating TADA

We benchmark TADA on 4 VALUE (Ziems et al., 2022, 2023) transformed versions of the GLUE Benchmark (Wang et al., 2018). As discussed in our limitations, these benchmarks are artificial but enable the evaluation of TADA across multiple tasks and dialects. First, we show how TADA compares to SAE models and task-specific baselines for African American English (AAE). Then, we show that TADA is effective across 4 global dialects of English. Finally, we perform an ablation to evaluate the contribution of each loss function.

For all TADA experiments, we train using 1,000 WiC sentences as described in Section 3.1. We train for 30 epochs with early stopping based on the lowest contrastive loss on a development set of 100 held-out WiC sentences. In Section 5, we report full hyperparameters along with the training details for SAE and VALUE models.

5 Training Details

TADA is trained with the ADAM optimizer for 30 epochs with batch size of 16 and with a hyperparameter search of 5e-4. We keep the model and epoch with lowest L_2 loss on the 100 held-out examples. Training takes approx. 30 minutes on an Nvidia GeForce RTX 2080 Ti.

To find this hyperparameter setup, we performed a grid search over batch sizes from 8, 16, 32 and learning rates from $5 \cdot 10^{-3}$, $5 \cdot 10^{-4}$, $5 \cdot 10^{-5}$ for AAVE and used the configuration with the lowest L_2 loss on the 100 held-out examples.

For all SAE and VALUE GLUE models, we finetune RoBERTa base for 10 epochs with the ADAM optimizer, a learning rate of $2 \cdot 10^{-5}$, a batch size of 16, and a linear learning rate warm-up of 6%. For all SAE and VALUE GLUE adapters, we finetune the original adapter architecture (Houlsby et al., 2019) inside RoBERTa base for 20 epochs with the ADAM optimizer, a learning rate of $1 \cdot 10^{-4}$, a batch size of 16, and a linear learning rate warm-up of 6%. Training all baseline models took approx. 3 days on an Nvidia GeForce RTX 2080 Ti. Additionally, we report experimental results on the BERT-base model in Appendix A1.

5.1 TADA vs. Task-Specific

Since ours is the first work to attempt task-agnostic dialect adaptation, we benchmark TADA in comparison to prior task-specific methods in Table 1.

We first establish pure SAE baselines for both full finetuning and adapter training (Houlsby et al., 2019). Interestingly, the gap between SAE performance and AAE performance is similar for adapters (-8.8) and full finetuning (-8.9) when trained on SAE. The minimal effects of the limited capacity of adapters on disparity indicate that dialectal discrepancy is largely within the pretrained LLM before finetuning. Without mitigation, SAE models alone perform poorly on non-SAE input.

We then train two task-specific dialect mitigation following the approach of VALUE, which augments training data with pseudo-dialect examples during finetuning. This is a strong baseline, as it allows the model to adapt specifically to in-domain augmented examples rather than the general sentences used to align TADA modules. When trained on augmented data, adapters (80.7 Avg.)² seem to outperform full finetuning (77.5 Avg.). We hypothesize that random initialization of adapters prevents

²Avg. refers to the mean performance across GLUE tasks.

conflicting gradients across dialects which can lead to negative transfer (Wang et al., 2020).

Finally, we combine TADA with task-specific SAE modules for our task-agnostic approach. TADA succeeds in our goal of generalizable performance improvements, yielding improved robustness for 6 out of 7 tasks for an average increase of 2.8 points on the GLUE benchmark. However, TADA performs 4% worse on average than task-specific VALUE-augmented adapters. These adapters are trained on larger amounts of dialectal training data directly from each task than TADA, which likely explains their superiority. However, as noted in the table these approaches scale training and storage linearly with the number of tasks, while TADA requires only a constant overhead.

These results are the first to indicate the possibility of task-agnostic dialect adaptation. While performance lags behind the task-specific intervention, these results indicate similar quality is possible with vastly improved scalability. This scalability across tasks is key to truly addressing dialect disparities as NLP has a growing impact across a larger number of tasks.

5.2 Cross-Dialectal Evaluation

We then confirm that TADA generalizes across regional dialects using 3 global dialect translations introduced from Ziems et al. (2023) in Table 2. Beyond AAE, we select Nigerian English and Indian English as they are each estimated to have over 100 million English speakers³, Singaporean English as it was identified as particularly challenging.

Despite not explicitly encoding any linguistic features, TADA is not dialect-agnostic. TADA improves average performance by +2.8, +0.3, +0.4, and +3.9 respectively for African American, Indian, Nigerian, and Singaporean Englishes.

Ultimately, this applicability across dialects reinforces TADAs potential as a general tool, but with key limitations at fully removing the dialect gap. Truly dialect-robust NLP requires generalization across both tasks and dialects, making measuring the performance of both essential. We recommend future works on dialect modeling evaluate both.

5.3 Ablation Study

Finally, we show the results from an ablation in Table 3 to evaluate the contributions of each loss

³Speaker estimates from the Oxford English Dictionary Introduction to Nigerian English and the Indian Census.

Test Dialect	CoLA		MNLI		QNLI		RTE		QQP		SST2		STS-B		Mean	
	Orig.	TADA														
SAE	58.3	87.2	87.2	87.2	93.2	93.2	70.8	70.8	93.9	93.9	90.5	90.5	90.5	90.5	83.5	83.5
AAVE	14.1	29.5	83.7	84.8	90.3	91.7	67.1	67.1	86.8	88.1	92.1	91.9	88.7	89.6	74.7	77.5 (+2.8)
Indian	16.4	15.0	82.6	83.6	89.1	90.3	66.8	66.8	86.4	87.0	90.9	91.1	88.5	88.9	74.4	74.7 (+0.3)
Nigerian	23.7	27.2	84.3	84.8	91.2	91.1	65.0	64.6	88.2	88.2	92.2	92.1	89.3	88.7	76.3	76.7 (+0.4)
Singaporean	-0.4	20.3	81.4	83.0	87.7	89.3	63.2	64.3	85.2	87.3	90.9	91.1	88.1	88.5	70.9	74.8 (+3.9)

Table 2: **Multi-Dialectal** evaluation results across all Tasks (Matthew’s Corr. for CoLA; Pearson-Spearman Corr. for STS-B; Accuracy for all others) for 4 Non-SAE Dialect Variants of GLUE created using Multi-VALUE.

Method	AAE Glue Performance							Mean
	COLA	MNLI	QNLI	RTE	QQP	SST2	STS-B	
TADA	29.5	84.8	91.7	67.1	88.1	91.9	89.6	77.5
$-L_{ms}$ (Eq. 3)	29.1	85.0	91.5	66.1	88.0	91.6	89.4	77.2 (-0.3)
$-L_{seq}$ (Eq. 1)	0.0	31.8	50.5	36.8	47.3	50.9	10.7	32.6 (-44.9)

Table 3: **TADA Loss Ablation** results for RoBERTa Base for the 7 GLUE Tasks (Matthew’s Corr. for CoLA; Pearson-Spearman Corr. for STS-B; Accuracy for all others) for African-American English. Our results show that the combined loss functions of TADA lead to the strongest results.

function to the final TADA methods. Contrastive loss alone yields close performance to TADA; it consistently underperforms the combined loss functions on 6 out of 7 tasks (-0.3 Avg.). This extends evidence for the efficacy of this simple loss function from the multilingual (Conneau et al., 2018) to the dialectal domain.

When contrastive loss is removed, the adversarial loss quickly becomes unstable and suffers from mode collapse. This leads to pathological results, with the resulting adapters harming performance for all tasks (-44.9 Avg.).

6 Conclusions

English dialects are underserved by NLP, but are both tractable targets for transfer learning and have huge speaking populations (Bird, 2022). Models which serve English speakers inherently serve a global population who use the language natively and as a second tongue.

However, current approaches to improve dialectal robustness in English have so far focused only on one task at a time. The scalability of these task-specific methods limits their impact as language technology applications become increasingly diverse and pervasive. We argue that task-agnostic dialectal methods are a clear, yet unexplored path to serve these communities effectively.

We propose a simple yet effective technique TADA to address this, utilizing morphosyntactic data augmentation and alignment loss at both the sequence and morphosyntactic level to train adapter modules. When composed with SAE task adapters,

TADA modules improve dialectal robustness consistently on the multi-task GLUE benchmark. Future work should work to further reduce the dialect discrepancy to create more inclusive and equitable English language technology.

Limitations

TADA makes use of the pseudo-dialectal translation systems of prior work Ziems et al. (2022, 2023). We rely on them as they are validated by dialect speakers and have been shown to be predictive of performance on Gold Dialect data. However, they were designed as stress tests of robustness which isolates morphology and syntax. We are therefore unsure how TADA performs when it faces the topical and register shifts which often are associated with naturally occurring dialects. These limitations are similar to localization issues in translated benchmarks (Moradshahi et al., 2020).

In this work, we evaluate TADA on only Encoder-only LLMs. Increasingly, both Encoder-Decoder and Decoder-only models are seeing wide-scale use due to their flexibility (Wang et al., 2022). Evaluating TADA and developing alternate tailored task-agnostic methodologies on these alternate LLM architectures is left to future work.

Ethics Statement

This work refers to linguist-drawn boundaries around dialects. However, dialects are not monolithic and are used in varied ways across sub-communities of speakers. Readers should there-

fore not understand TADA to remove discrepancies across all speakers as improvements may vary within subcommunities within a dialect (Koenecke et al., 2020). Additionally, as TADA is task-agnostic, it is especially vulnerable to dual use. To mitigate this, we will release TADA under a license that forbids usage with intent to deceive, discriminate, harass or surveil dialect-speaking communities in a targeted fashion.

Acknowledgements

We are thankful to Yanzhe Zhang and the anonymous ACL reviewers for their helpful feedback.

References

- Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. [Cross-lingual word embeddings for low-resource language modeling](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 937–947, Valencia, Spain. Association for Computational Linguistics.
- Anmol Agarwal, Jigar Gupta, Rahul Goel, Shyam Upadhyay, Pankaj Joshi, and Rengarajan Aravamudhan. 2022. Cst5: Data augmentation for code-switched semantic parsing. *arXiv preprint arXiv:2211.07514*.
- Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. 2021. [The low-resource double bind: An empirical study of pruning for low-resource machine translation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3316–3333, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. [A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.
- Steven Bird. 2022. [Local languages, third spaces, and other high-resource scenarios](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.
- Terra Blevins, Robert Kwiatkowski, Jamie MacBeth, Kathleen McKeown, Desmond Patton, and Owen Rambow. 2016. [Automatically processing tweets from gang-involved youth: Towards detecting loss and aggression](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2196–2206, Osaka, Japan. The COLING 2016 Organizing Committee.
- Su Lin Blodgett, Johnny Wei, and Brendan O’Connor. 2018. [Twitter Universal Dependency parsing for African-American and mainstream American English](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Melbourne, Australia. Association for Computational Linguistics.
- Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladha, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Porteance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. [On the opportunities and risks of foundation models](#).
- Shuguang Chen, Gustavo Aguilar, Anirudh Srinivasan, Mona Diab, and Thamar Solorio. 2022. Cales 2021 shared task: Machine translation for code-switched data. *arXiv preprint arXiv:2202.09625*.
- 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.
- Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. **XNLI: Evaluating cross-lingual sentence representations**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.
- Marta R. Costa-jussà. 2017. **Why Catalan-Spanish neural machine translation? analysis, comparison and combination with standard rule and phrase-based technologies**. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 55–62, Valencia, Spain. Association for Computational Linguistics.
- Marta R. Costa-jussà, Marcos Zampieri, and Santanu Pal. 2018. **A neural approach to language variety translation**. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 275–282, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Dorottya Demszky, Devyani Sharma, Jonathan Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021. **Learning to recognize dialect features**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2315–2338, Online. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. **BERT: Pre-training of deep bidirectional transformers for language understanding**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Mona Diab, Nizar Habash, Owen Rambow, Mohamed Altantawy, and Yassine Benajiba. 2010. Colaba: Arabic dialect annotation and processing. In *Lrec workshop on semitic language processing*, pages 66–74. Citeseer.
- Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. **Learning crosslingual word embeddings without bilingual corpora**. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1285–1295, Austin, Texas. Association for Computational Linguistics.
- Kawin Ethayarajh and Dan Jurafsky. 2020. **Utility is in the eye of the user: A critique of NLP leaderboards**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.
- Edouard Grave, Armand Joulin, and Quentin Berthet. 2019. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR.
- Suchin Gururangan, Dallas Card, Sarah K. Dreier, Emily K. Gade, Leroy Z. Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. 2022. **Whose language counts as high quality? measuring language ideologies in text data selection**.
- Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. 2022. **ArzEn-ST: A three-way speech translation corpus for code-switched Egyptian Arabic-English**. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 119–130, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.
- Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2016. **Learning a POS tagger for AAVE-like language**. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1115–1120, San Diego, California. Association for Computational Linguistics.
- David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. **Incorporating dialectal variability for socially equitable language identification**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57, Vancouver, Canada. Association for Computational Linguistics.
- Filip Klubička, Gema Ramírez-Sánchez, and Nikola Ljubešić. 2016. **Collaborative development of a rule-based machine translator between Croatian and Serbian**. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 361–367.
- Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

- Guillaume Lample, Alexis Conneau, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.
- Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2020. Exploring versatile generative language model via parameter-efficient transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 441–459, Online. Association for Computational Linguistics.
- 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*.
- George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.
- Mehrad Moradshahi, Giovanni Campagna, Sina Semnani, Silei Xu, and Monica Lam. 2020. Localizing open-ontology QA semantic parsers in a day using machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5970–5983, Online. Association for Computational Linguistics.
- Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.
- Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.
- Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.
- Maja Popović, Alberto Poncelas, Marija Brkić, and Andy Way. 2020. Neural machine translation for translating into croatian and serbian. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 102–113.
- Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer nlp. *arXiv preprint arXiv:2205.12586*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.
- Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Wael Salloum and Nizar Habash. 2013. Dialectal Arabic to English machine translation: Pivoting through Modern Standard Arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 348–358, Atlanta, Georgia. Association for Computational Linguistics.
- Karin Kipper Schuler. 2005. VerbNet: A broad-coverage, comprehensive verb lexicon. University of Pennsylvania.
- Asa Cooper Stickland and Iain Murray. 2019. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. In *International Conference on Machine Learning*, pages 5986–5995. PMLR.
- Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176.
- 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.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *BlackboxNLP@EMNLP*.
- Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, and Colin Raffel. 2022. What language model architecture and pretraining objective work best for zero-shot generalization? *arXiv preprint arXiv:2204.05832*.
- Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*,

pages 4438–4450, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. 2018. [Grounding interactive machine learning tool design in how non-experts actually build models](#). In *Proceedings of the 2018 Designing Interactive Systems Conference*, DIS ’18, page 573–584, New York, NY, USA. Association for Computing Machinery.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. [Earth mover’s distance minimization for unsupervised bilingual lexicon induction](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. [VALUE: Understanding dialect disparity in NLU](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3701–3720, Dublin, Ireland. Association for Computational Linguistics.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-[VALUE](#): A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.

Dialect Adaptation Details				AAE Glue Performance							
Approach	Method	Task-Agnostic	Dialect Params.	COLA	MNLI	QNLI	RTE	QQP	SST2	STS-B	Mean
N/A	Finetuning	✓	0	36.0	79.6	89.2	65.3	86.2	89.7	87.4	76.2
N/A	Adapters	✓	0	31.4	80.8	89.2	62.1	86.0	89.8	86.9	75.1
VALUE	Finetuning	✗	$T \times 110M$	36.2	83.0	89.7	61.4	88.6	89.6	88.2	76.7
VALUE	Adapters	✗	$T \times 895K$	36.3	82.0	89.5	66.8	85.6	88.8	88.5	76.8
TADA	Adapters	✓	$895K$	38.3	81.5	89.0	62.1	87.0	90.0	88.0	76.6

Table A1: **Dialect Adaptation GLUE** results of BERT Base (Devlin et al., 2019) for the 7 GLUE Tasks (Matthew’s Corr. for CoLA; Pearson-Spearman Corr. for STS-B; Accuracy for all others). T is the number of target tasks for dialect adaptation.