Dynamic Aggregation and Auto-Discovery of Linguistic Features

Yanchen Liu

Stanford | NLP

DADA: Dialect Adaptation via Dynamic Aggregation of Linguistic Rules

Yanchen Liu, William Held, Diyi Yang

will appear on EMNLP 2023

Non-Standard Linguistic Features

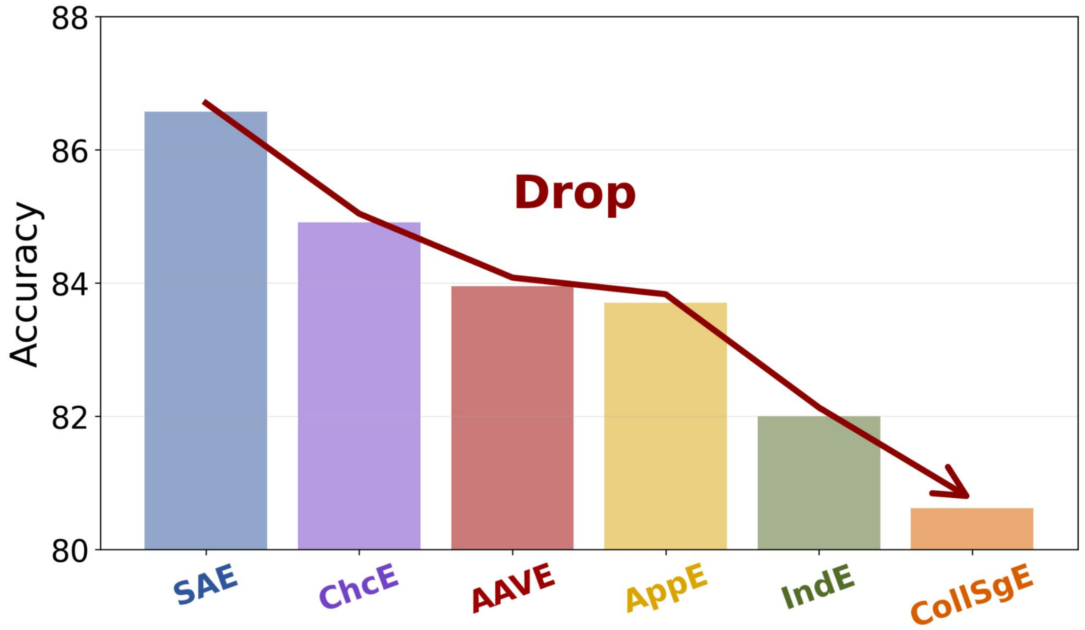
Stanford NLP

- ❑ language usage that deviates from the conventions
- ❑ often associated with specific social or cultural groups

Dialect: A Group of Non-Standard Linguistic Variations in a Language

Fails on Dialects

- ❑ Existing models mainly focus on Standard American English (**SAE**)
- ❑ Significant **performance drop↓** when applied to English Dialects



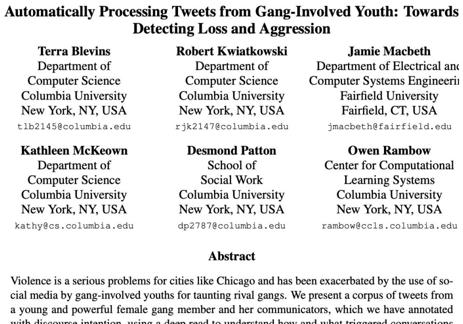
Previous Work on Dialect Adaptation

Stanford NLP

Mainly focused on targeted adaptation to **a specific dialect**

- Human Annotation (Blevins et al., 2016, Blodgett et al., 2018)
- Weak Supervision (Jorgensen et al. 2016, Jurgens et al. 2017)
- Alignment (TADA 2023)

⚠️ **Highly accurate dialect identification systems are required!**

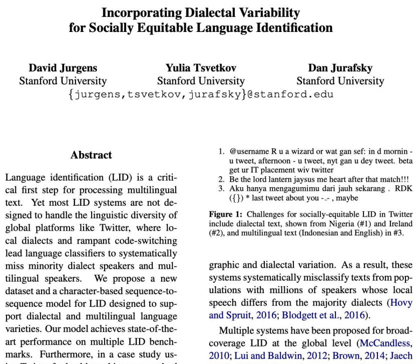


Automatically Processing Tweets from Gang-Involved Youth: Towards Detecting Loss and Aggression

Terra Blevins, Department of Computer Science, Columbia University, New York, NY, USA, tlb2145@columbia.edu
Robert Kwiatkowski, Department of Computer Science, Columbia University, New York, NY, USA, rjk2147@columbia.edu
Jamie Macbeth, Department of Electrical and Computer Systems Engineering, Fairfield University, Fairfield, CT, USA, jmacbeth@fairfield.edu
Kathleen McKeown, Department of Computer Science, Columbia University, New York, NY, USA, kathy@cs.columbia.edu
Desmond Patton, School of Social Work, Columbia University, New York, NY, USA, dp2787@columbia.edu
Owen Rambow, Center for Computational Learning Systems, Columbia University, New York, NY, USA, rambow@ccls.columbia.edu

Abstract

Violence is a serious problems for cities like Chicago and has been exacerbated by the use of social media by gang-involved youths for taunting rival gangs. We present a corpus of tweets from a young and powerful female gang member and her communicators, which we have annotated



Incorporating Dialectal Variability for Socially Equitable Language Identification

David Jurgens, Stanford University
Yulia Tsvetkov, Stanford University
Dan Jurafsky, Stanford University
{jurgens,tsvetkov,jurafsky}@stanford.edu

Abstract

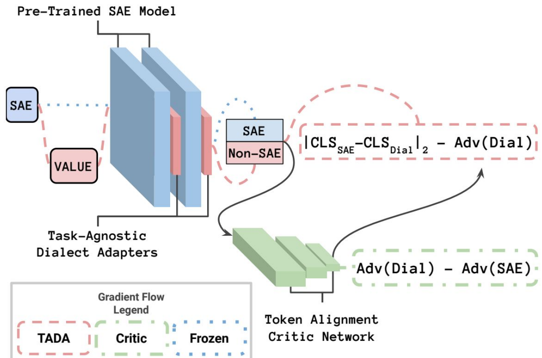
Language identification (LID) is a critical first step for processing multilingual text. Yet most LID systems are not designed to handle the linguistic diversity of global platforms like Twitter, where local dialects and rampant code-switching lead language classifiers to systematically miss minority dialect speakers and multilingual speakers. We propose a new dataset and a character-based sequence-to-sequence model for LID designed to support dialectal and multilingual language varieties. Our model achieves state-of-the-art performance on multiple LID benchmarks. Furthermore, in a case study us-

1. @username R u a wizard or wat gan sef: in d mornin - u tweet, afternoon - u tweet, nyt gan u dey tweet. beta get ur IT placement wiv twitter
2. Be the lord lantern jaysus me heart after that match!!!
3. Aku hanya mengagumimu dari jauh sekarang . RDK ({}) * last tweet about you -.- , maybe

Figure 1: Challenges for socially-equitable LID in Twitter include dialectal text, shown from Nigeria (#1) and Ireland (#2), and multilingual text (Indonesian and English) in #3.

graphic and dialectal variation. As a result, these systems systematically misclassify texts from populations with millions of speakers whose local speech differs from the majority dialects (Hovy and Spruit, 2016; Blodgett et al., 2016).

Multiple systems have been proposed for broad-coverage LID at the global level (McCandless, 2010; Lui and Baldwin, 2012; Brown, 2014; Jaech



However!!! Inherent **Flexibility** of Dialects

Stanford NLP

- ❑ Flexible Boundaries

  => no highly accurate dialect identification systems available

- ❑ Vary Depending on Personal and Social Contexts

  => dialects do not neatly fit into predefined categories

🌟 **Accommodate the diversity of dialects from a Fine-Grained perspective Linguistic Features**

Method

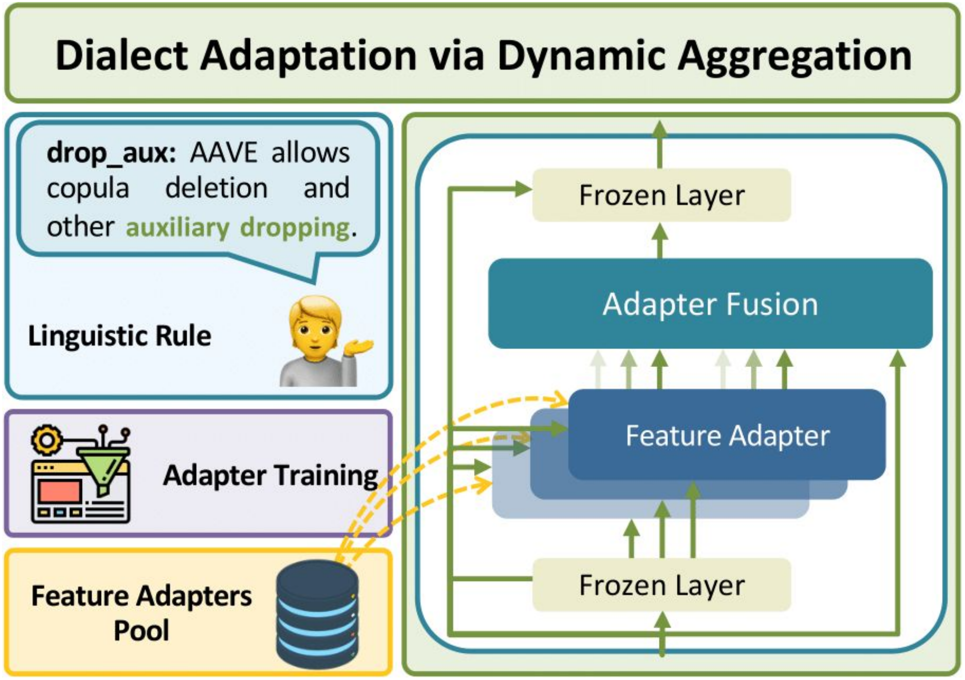
Dialect Adaptation via Dynamic Aggregation

🎖️ Modular and Dynamic

- ❑ Multi-Dialectal Robustness
  - ❑ Input Dialect-Agnostic
  - ❑ Personal- and Social-Contextual

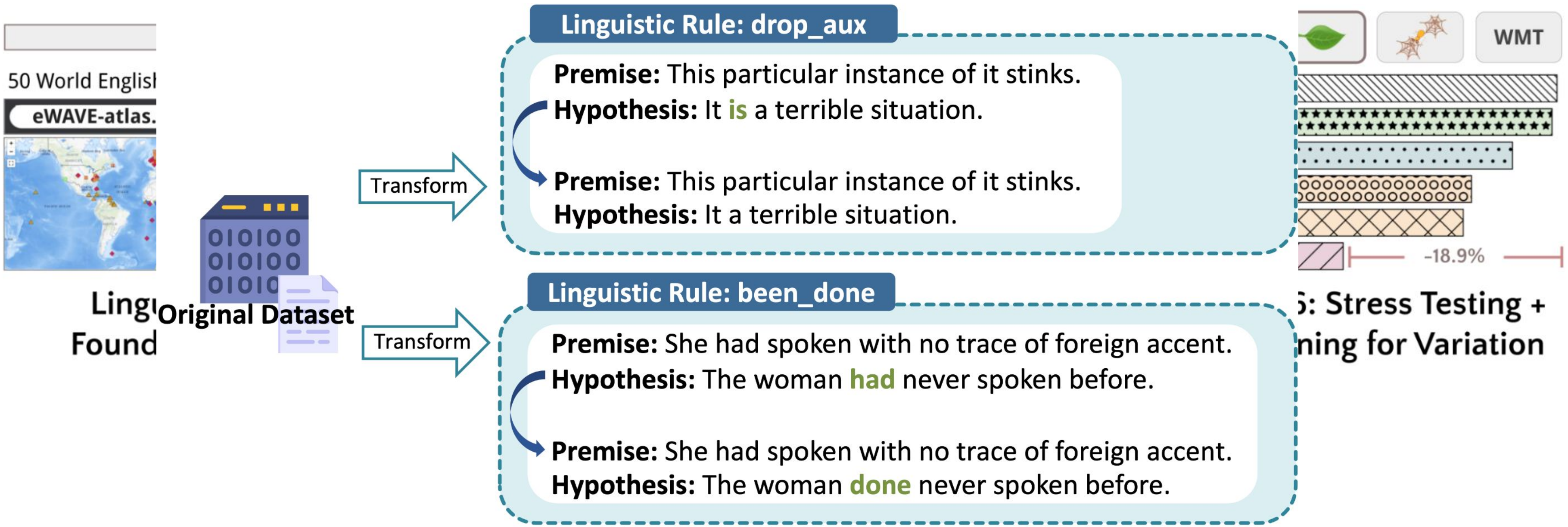
🌟 Within Only 3 Steps

1. Synthetic Datasets Construction
2. Feature Adapter Training
3. Dynamic Aggregation



Step 1: Synthetic Datasets Construction

🔖 **Construct a transformed dataset for each non-standard (morphosyntactic) linguistic feature.**

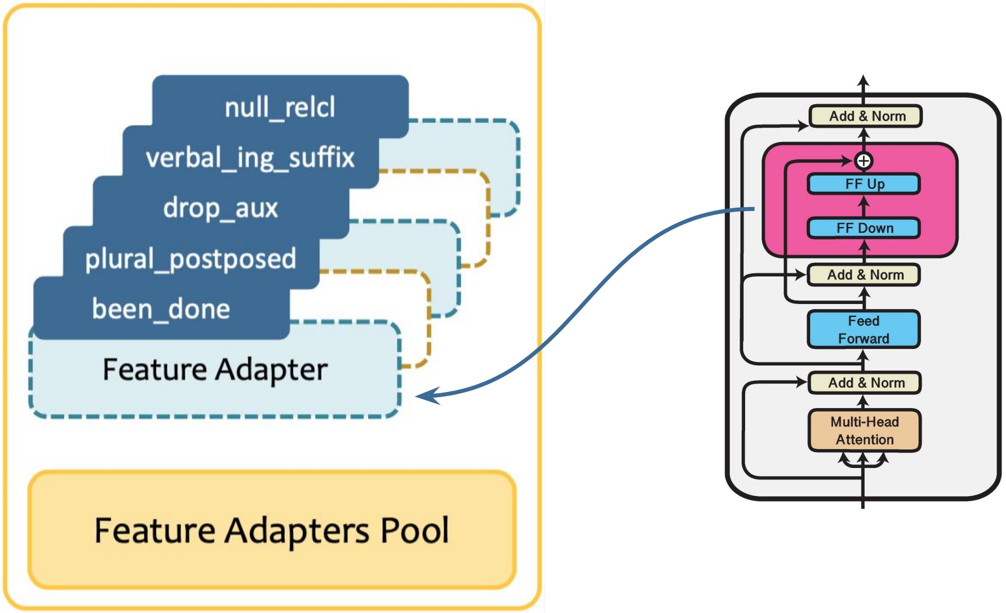


[2] Ziems et al. 2022. Multi-VALUE: Evaluating Cross-dialectal NLP. ACL 2023.

Step 2: Feature Adapter Training

Stanford NLP

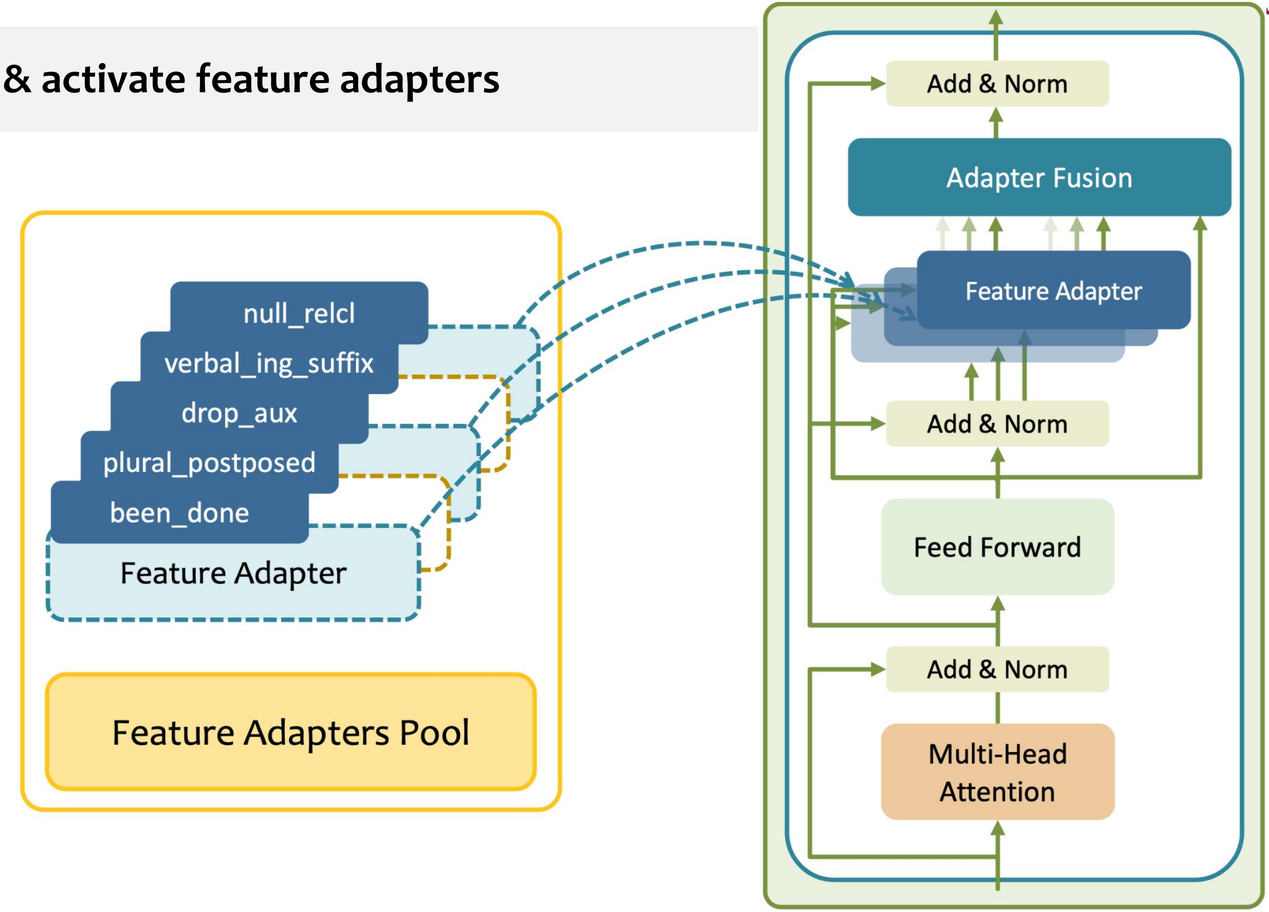
Train a feature adapter for each non-standard linguistic feature.



Step 3: Dynamic Aggregation

Stanford NLP

🔖 **Aggregate & activate feature adapters dynamically.**

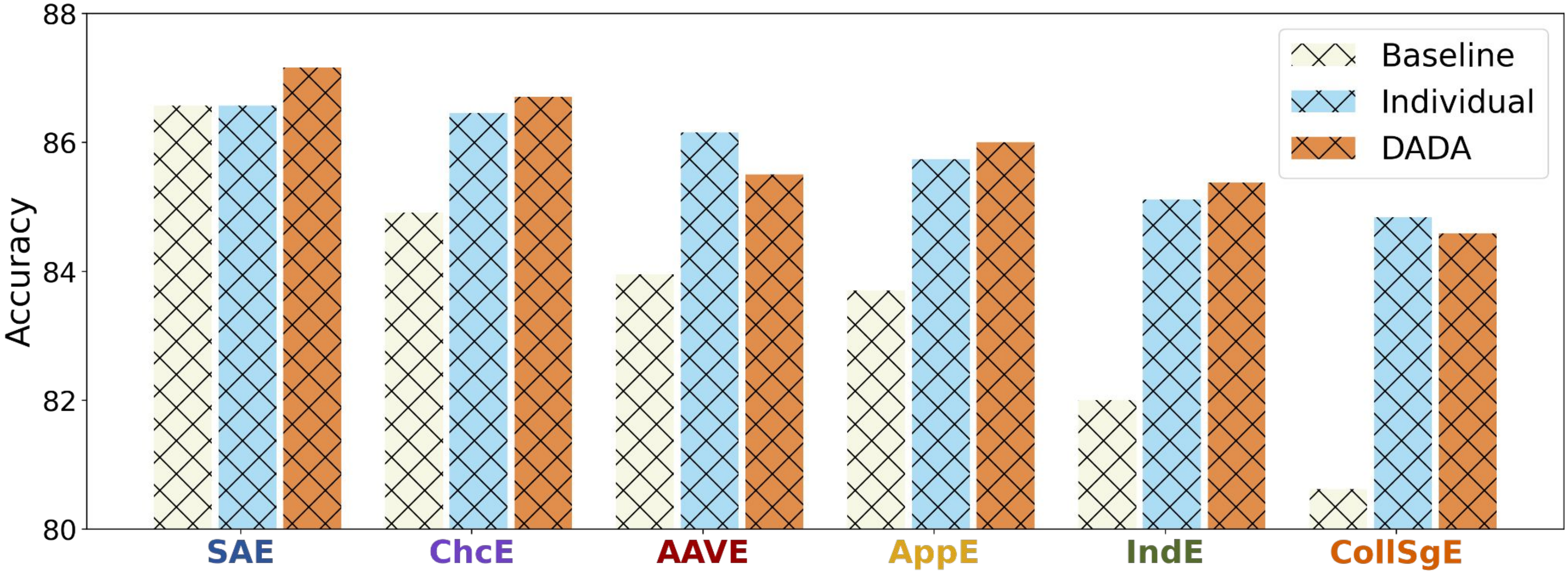


Experiments

1/ DADA Can Improve Multi-Dialectal Robustness

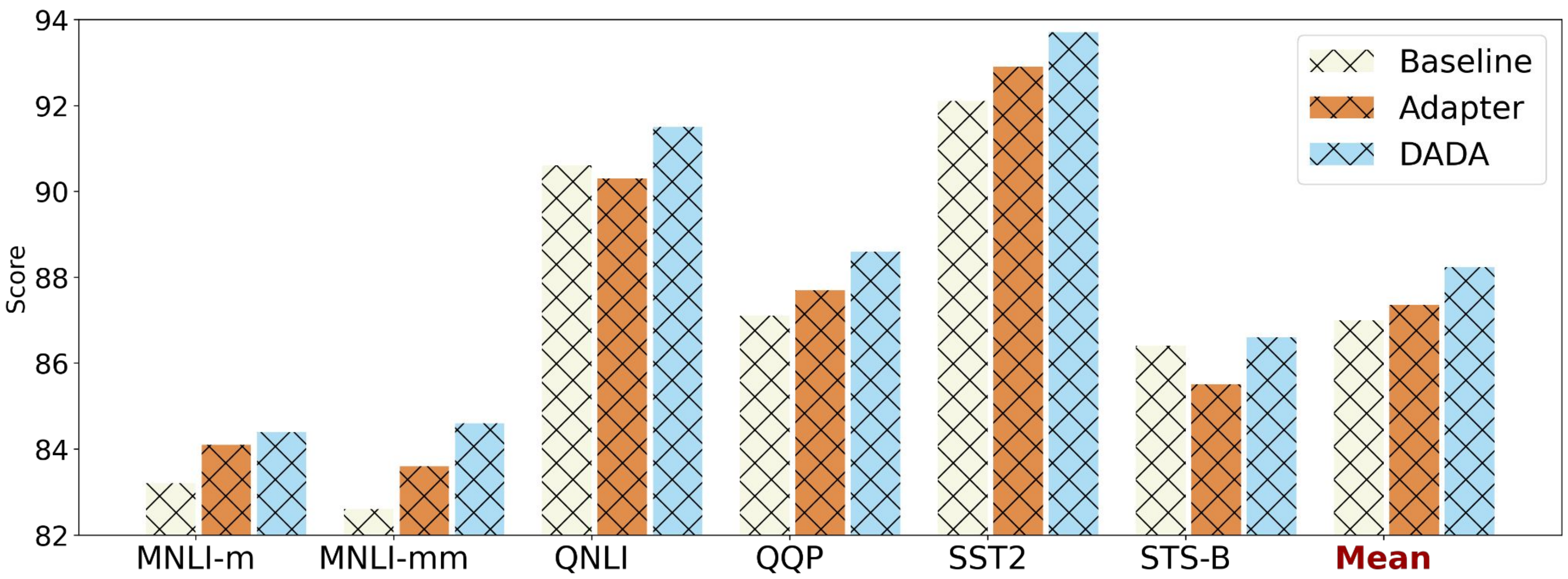
Stanford NLP

Adapt **SAE** model to multiple dialect variants simultaneously: **AppE**, **ChcE**, **CollSgE**, **IndE**, **AAVE**



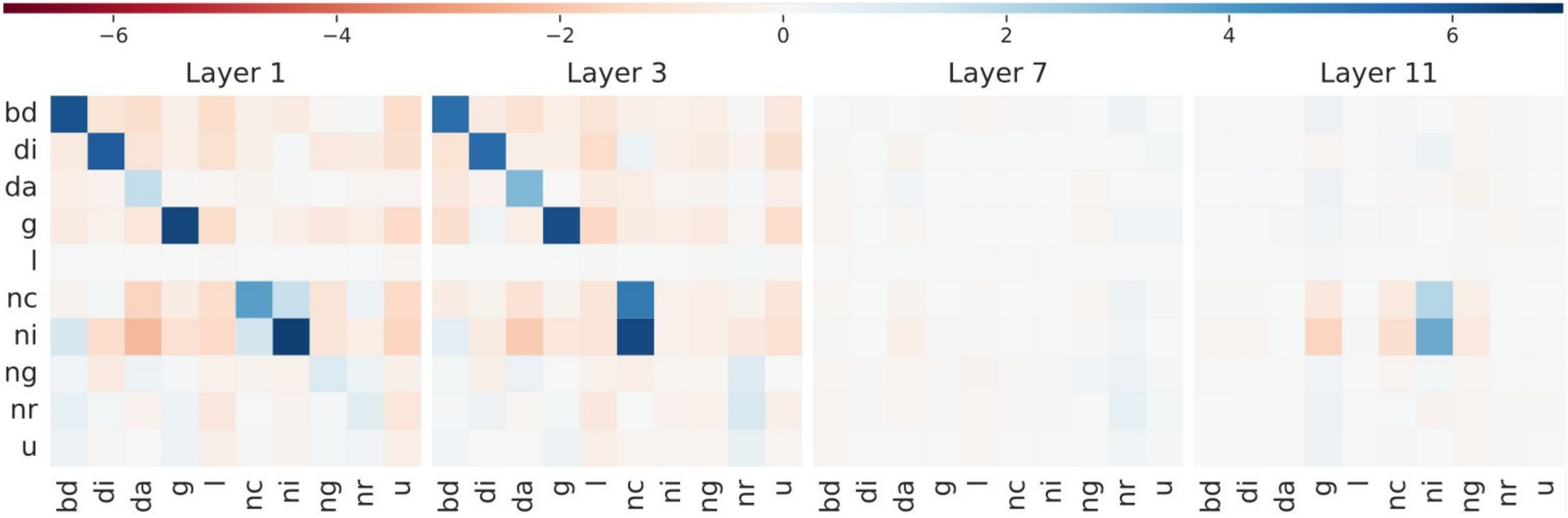
2/ DADA Can Be Task-Agnostic

Adapt instruction-tuned **SAE** model to the dialect variants for multiple tasks



3/ DADA Has Great Interpretability!

Correlation Coefficients for **AAVE** Adaptation



We use abbreviations for certain terms, such as "nc" for "negative concord."

Conclusion and Future Work

Dynamic Aggregation of Linguistic Features

- A **Fine-grained** and **Modular** Method for Dialect Adaptation
  - Improve **Multi-Dialect** and **Multi-Task** Robustness
    - No need for highly accurate dialect identification systems
    - Taking personal and social context into account
    - Applicable to task-agnostic instruction-tuned LLMs
  - **Interpretability**, reusability and extensibility

But!!!

Non-Standard Linguistic Features

- ❑ are curated by linguists (eWAVE) and
- ❑ play a crucial role in a wide range of applications.

However, the manual curation of linguistic rules can be **expensive** and **expertise-intensive**.

Empower Linguistic Research with LLMs

Large Language Models Can Discover Linguistic Features

(ongoing)