Towards Equitable Natural Language Understanding Systems for Dialectal Cohorts: Debiasing Training Data

Khadige Abboud and Gokmen Oz
Alexa AI, Amazon

LREC-COLING May 2024

amazon | science

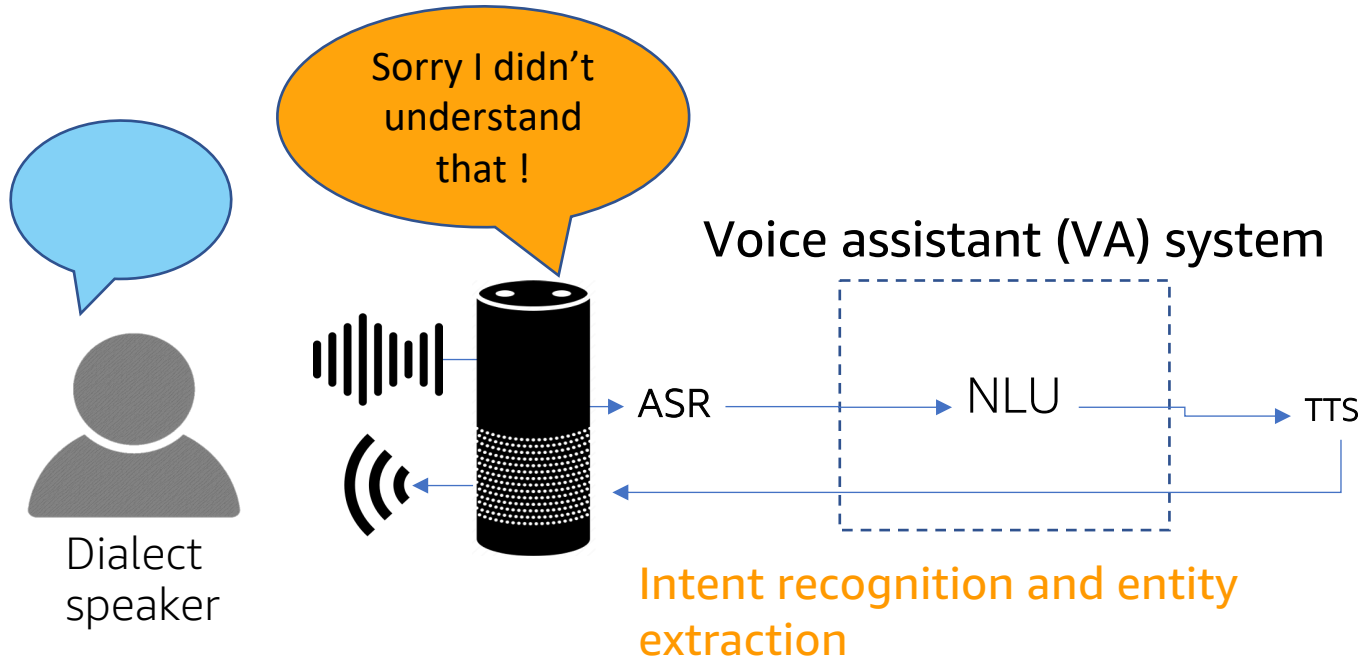
Background

- Training natural language understanding (NLU) systems relies primarily on standard language resources
  - Difficult and costly to secure large amounts of high-quality dialectal data

- Biases exit in LLMs against low-resource languages or dialects and it can start as early as
  - the tokenizer level
  - data quality filters that are applied to data sources prior to the model training

Dialectal bias push certain groups away from a technology

In households, the common location for VAs, dialectal language is more likely to be used



→ Leading to biased and inequitable NLU technology that underserves dialectal speakers.

Challenges in dialectal variants

Different vocabulary and morphology

“Thank you” in German

danke (de) → merci (gsw)

“How are you?” in Arabic

(ar) كيف حالك؟/kaifa halok/ → (EG) ازيك ؟/izayak/ (GL) شلونك/shlonik/

→ Differences vary by social, cultural and economic factors

Challenges in dialectal variants

Orthographic inconsistencies

- Dialects are **spoken languages** and not standardized written languages
- Lack of writing standards causes
  - Inconsistencies in transcribing data

slowly

/a’mahlak/عمهلك vs على مهلك/ala mahlak/

→ Further amplifies the scarcity of already low-resource language

Challenges in dialectal variants

Entity dialectal variants for NLU to capture and learn

can you turn on the lights

Modern Standard Arabic (ar): هل يمكنك فتح الإضاءة

Gulf Arabic (GL): تقدرين تشبّين الليتات

Egyptian Arabic (EG): ممكن تولّعي النور

→ Amplify data needs

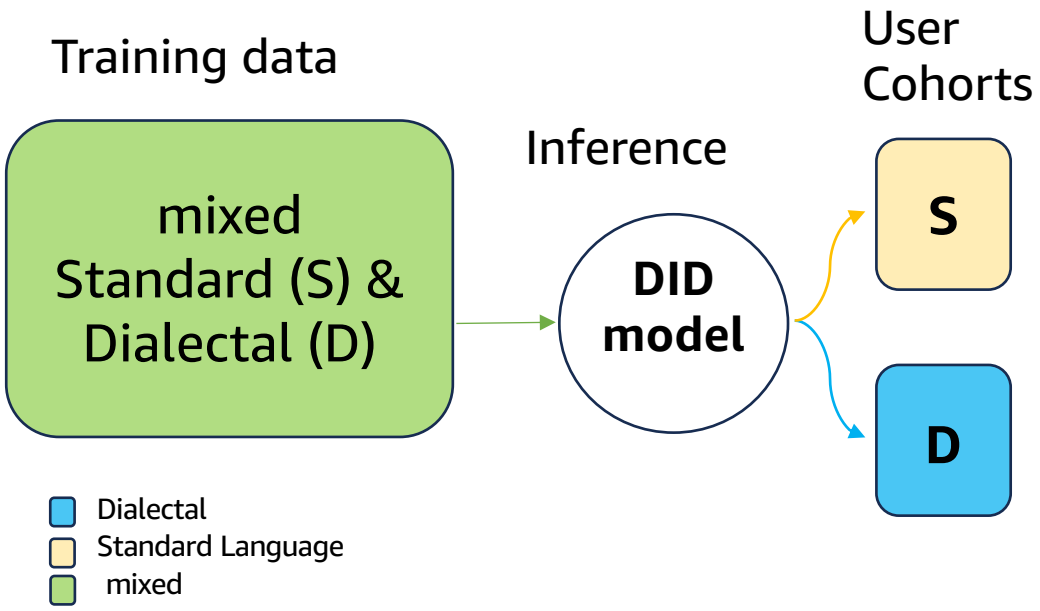
Entity extraction across dialectal varieties

Goal

- Dialect-based debiasing of training data by targeted data-augmentation to boost NLU performance for dialectal cohorts in **voice assistant (VA) systems**
- **Challenge:** The dialectal makeup of traffic in real VA systems and the information about NLU performance on dialectal cohorts are not readily available.

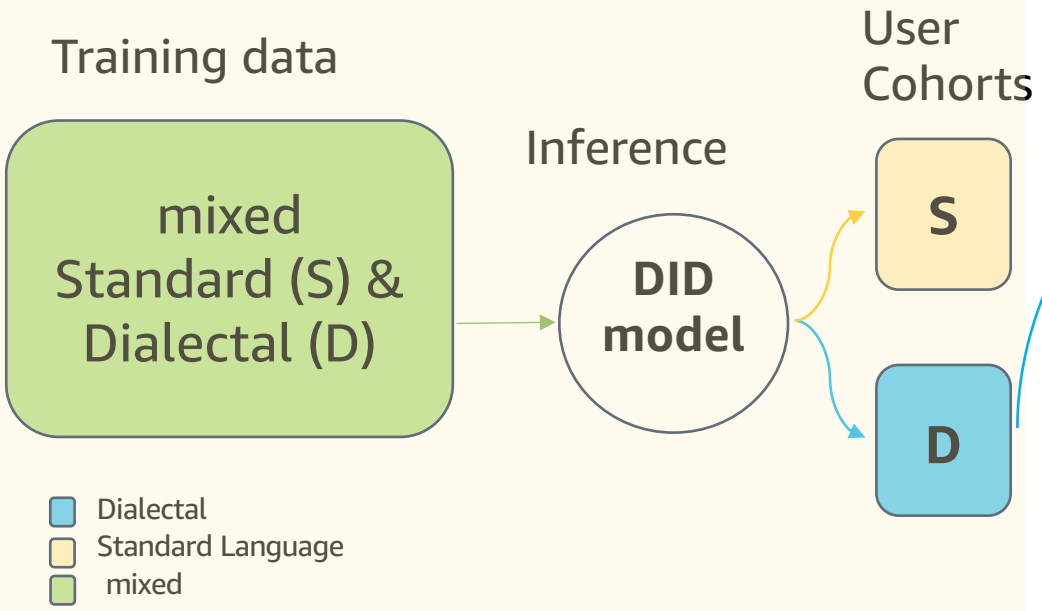
Extracting dialectal sub-population from training data

Extracting dialectal cohorts from training data

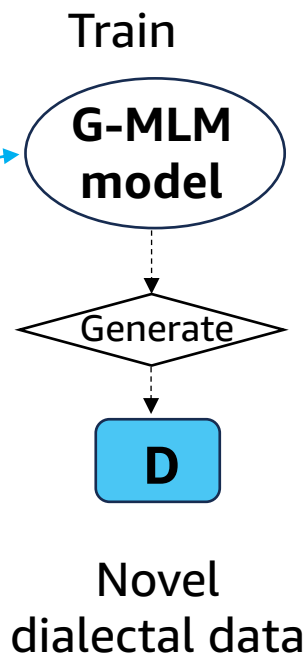


Dialect-data augmentation

Extracting dialectal cohorts from training data



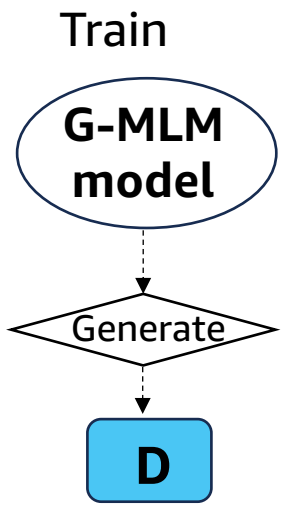
Augmenting dialectal data



+ Le, Thu, et al. "De-biasing training data distribution using targeted data enrichment techniques." DLP workshop, KDD 2022

Dialect-data augmentation

Augmenting dialectal data



play song by **nickelback** [sep] Music PlayMusic Other MediaType Other Artist

play song by [MASK] [sep] Music PlayMusic Other MediaType Other Artist

play song by michael jackson [sep] Music PlayMusic Other MediaType Other Artist

play song by adele [sep] Music PlayMusic Other MediaType Other Artist

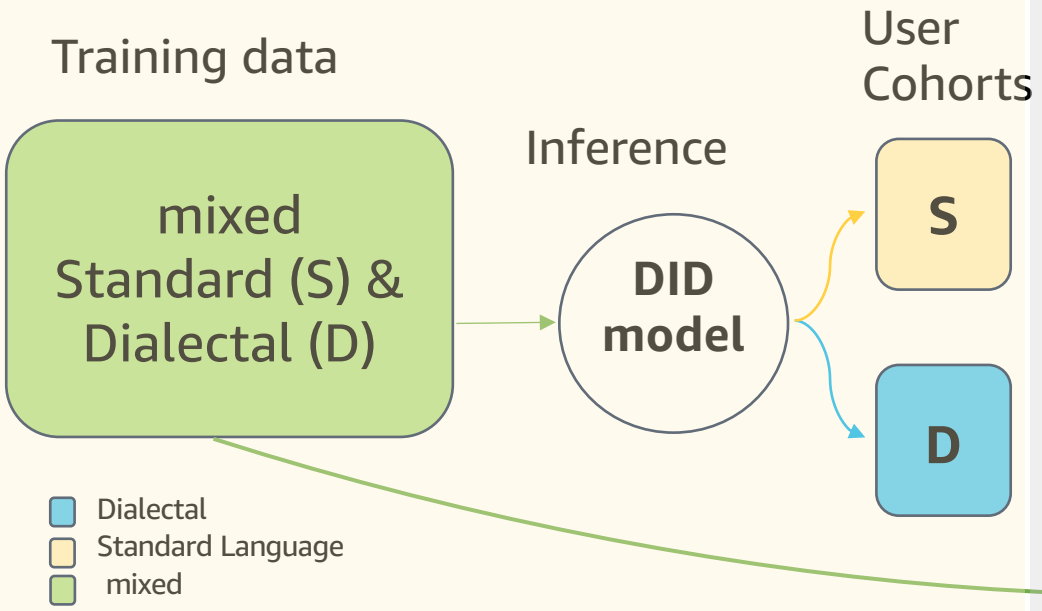
→ variable masking probability = measures how *replaceable* the word is, calculated as the number of times pairs of utterances in the seed intent differ only on this word.

For finetuning the MLM generator we use a pretrained monolingual BERT base models for Arabic and German, because we want the model to generate utterances in our target language.

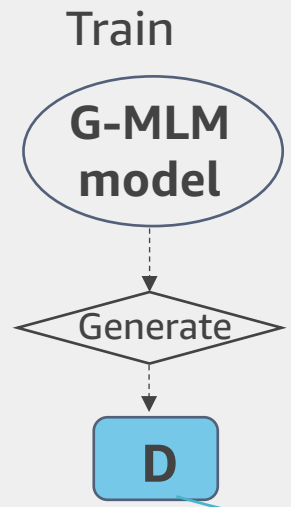
+ Le, Thu, et al. "De-biasing training data distribution using targeted data enrichment techniques." DLP workshop, KDD 2022

Finetune dialect-debiased NLU model

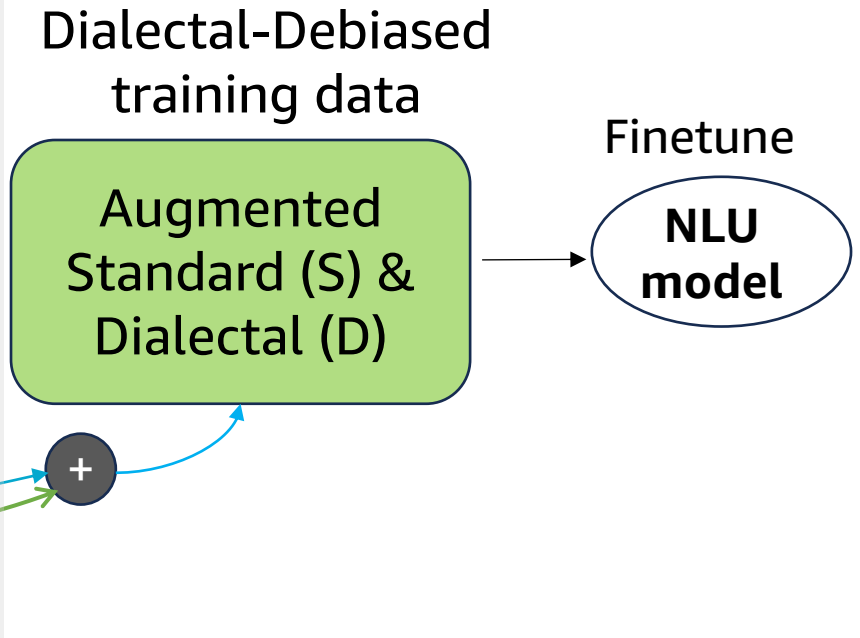
Extracting dialectal cohorts from training data



Augmenting dialectal data



Feed generated data to finetune NLU model



Data - Dialect tagged data for dialect identification (DID)

Training two monolingual DID

- SwissDial* 26 hours of studio-quality recordings of 8 Swiss dialects in addition to standard German (de)
- mTurk: internally collected for 2 Arabic varieties Egyptian (EG) and Gulf (GL) - Saudi and Emirati in addition to the standard Arabic (ar)

Evaluating DID model

Language	Standard	Dialect
ar	70	77.5
de	87.6	90.4

Dataset	Language	Size (utterance)	
		Train	Test
SwissDial*	de + dialects	42,134	-
xSID†	de	-	500
	de-gsw	-	500
	de-st	-	500
ar mTurk	ar + dialects	413,459	-
MADAR•	ar-MSA	-	200
	ar-EG	-	600
	ar-GL	-	600

e.g. حطي أغنية بتاعت nickleback ar-EG

*Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2019. SwissDial: Parallel multidialectal corpus of spoken swiss german. ArXiv, abs/1910.01108.

† Van Der Goot et al., "From masked language modeling to translation: Non- English auxiliary tasks improve zero-shot spoken language understanding. ACL pages 2479– 2497

• Bouamor, Houda et al. 2018. The MADAR Arabic dialect corpus and lexicon. LREC 2018

Data - Annotated data for NLU

xSID[2] benchmark

- Standard German and St. Galler-Dütsch (de-gsw) and South Tyrolean (de-st) dialects: 8 domains
- For training mix dialectal + standard de to emulate our use-case

de/ar traffic from VA commercial system

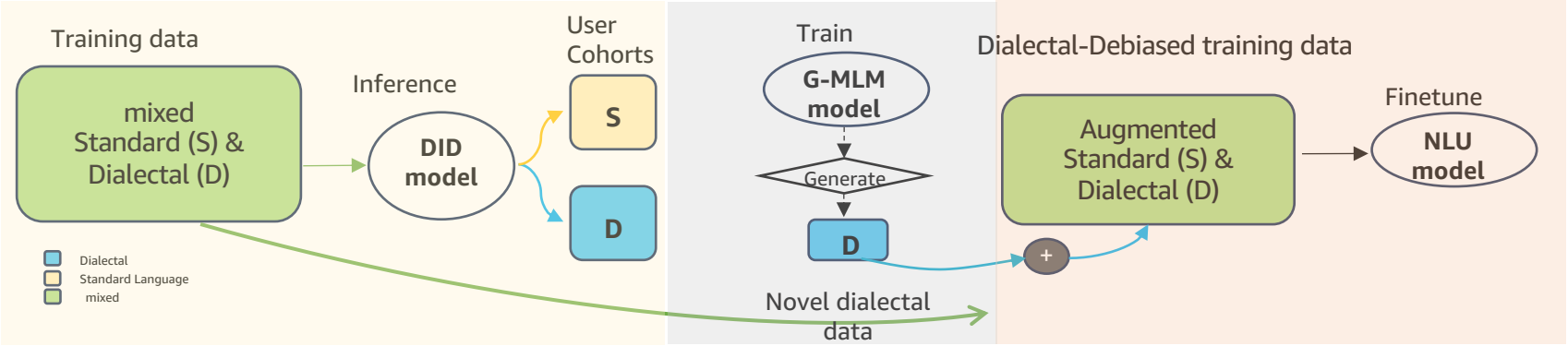
- mixed german/Arabic data spaning 22 domains
- Dialectal makeup unknown

Dataset	Language	Size (utterance)	
		Train	Test
xSID†	de + dialects	44,507	-
	de	-	500
	de-gsw	-	500
	de-st	-	500
de traffic	de + dialects	16,301,859	1,260,032
ar traffic	ar + dialects	1,423,139	144,533
ar mTurk	ar-MSA	-	24,810
	ar-GL	-	21,298
	ar-EG	-	3,594

e.g., Music, PlayMusic,
حطي|Other أغنية|MediaType بتاعت|Other nickleback|Artist

† Van Der Goot et al., “From masked language modeling to translation: Non- English auxiliary tasks improve zero-shot spoken language understanding. ACL pages 2479– 2497

Experiments



Pretrained multilingual LM is finetuned for NLU tasks

- BASELINE: finetuned on original data with unknown dialectal makeup
- Dialect debiased model: finetuned on original data + dialectal augmented data

Baseline underperforms on dialectal cohorts

- Using our DID model we can extract dialectal subsets of the evaluation data

- Results expose baseline performance disparity for dialectal cohorts compared standard-language speakers.

	Dataset		
Method	xSID (de)	de	ar
Average	14.26	8.74	-14.53
Overall	0	20.56	14.51

Results: Relative Semantic error rate (SemER) difference (% change) between baseline and debiased model, on the xSID† German dataset

Domain	xSID German de-standard	de-dialect
Alarm	-13.48	-20.65
Books	10.13	-12.01
CreativeWorks	2.87	-11.26
Events	3.76	3.52
Music	-11.58	-26.95
Reminder	2.40	-14.10
Reservations	7.37	-15.13
Weather	1.03	-14.04
Overall	**-1.11**	**-14.87**

Results: SemER difference (%) between baseline and debiased model the Arabic (ar) and German (de) large-scale VA commercial datasets.

	VA-system Arabic		VA-system German	
Domain	**ar-standard**	**ar-dialect**	**de-standard**	**de-dialect**
Knowledge	-12.0	-6.8	-2.39	-11.52
Events	-7.14	-3.58	-0.20	9.90
Communication	-0.3	-3.3	5.48	-4.87
SmartHome	1.3	-0.3	1.57	1.50
Music	2.9	2.0	-0.65	-0.20
Notifications	-7.3	-1.3	-4.87	1.20
Weather	27.7	1.4	-7.94	-6.38
Overall	**-1.04**	**-1.53**	0.05	**-1.32**

Improvements are not driven by training data volume

Comparing dialect-based debiasing with random upsampling

	Dataset		
Model	de-dialect (xSID)	de-dialect	ar-dialect
Random upsampling	-6.76	-0.96	16.57
Dialect-based debias	**-14.87**	**-1.32**	**-1.53**

Performance boost still exists on independently annotated testset

Assumption: the use of the same DID model for extracting dialectal cohorts from training and evaluation datasets

Evaluating **ar** model on an independently annotated testset with dialect tags

	Test subset		
Annotation method	All	ar-standard	ar-dialect
Human annotated	-0.56	-0.12	-1.02
DID model annotated	-0.94	-1.04	-1.53

Dialect-based debiasing improves larger models

- Question: Is the proposed method is effective for LLMs beyond BERT-like models?

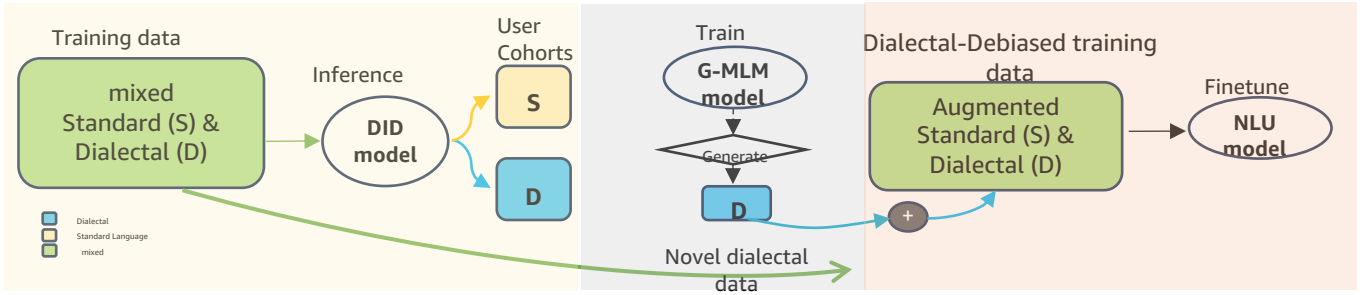
  Finetune LLMs with decoder-encoder and decoder-only seq2seq architectures of different sizes to generate labeled utterances

SemER performance on dialectal testset of xSID[†] German dataset for different LLMs.

Model size	baseline	debiased	%(change)
5B	0.41	0.33	**-19.51%**
7B	0.31	0.27	**-12.90%**
20B	0.34	0.32	**-5.88%**
30B	0.24	0.23	**-4.17%**

Conclusions

Dialect debiasing framework



- Dialect-based debiasing reduces dialectal disparity when tested on **two languages with high dialectal richness**, **Arabic and German**. Both languages have large populations of speakers and **exhibit significant linguistic diversity,** including differences in vocabulary and text.
- Proposed framework is effective on both open-source and large-scale VA commercial datasets.
- Ablation studies show that boost in performance on dialectal cohorts is **not** driven by volume only, is seen on annotated **data independent of DID model**, and is effective on LLM seq2seq architectures