MaiNLP

LMU Munich

EKATERINA ARTEMOVA

VERENA BLASCHKE

BARBARA PLANK

Exploring the Robustness of Task-oriented Dialogue Systems for Colloquial German Varieties

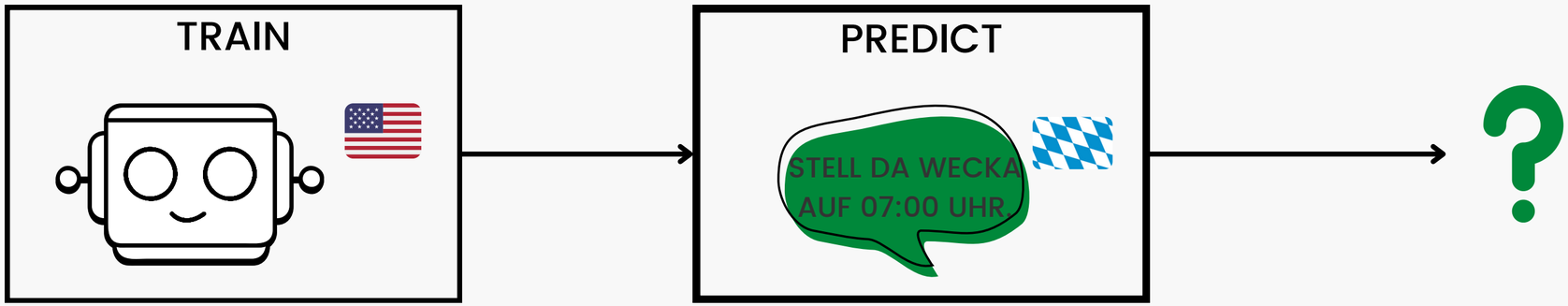
github.com/mainlp/dialect-ToD-robustness

I Motivation

- Cross-lingual ToD systems train a single model in English for intent recognition and slot-filling, applying it zero-shot to other languages;

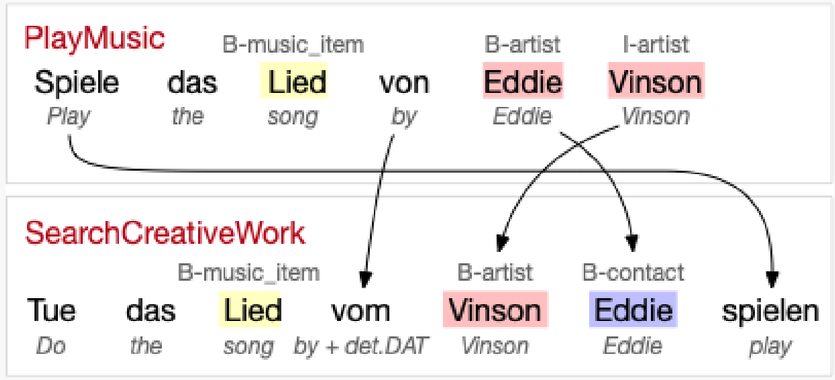
I Motivation

- Cross-lingual ToD systems train a single model in English for intent recognition and slot-filling, applying it zero-shot to other languages;
- However, they often overlook transfer to lower-resource colloquial varieties due to limited test data.

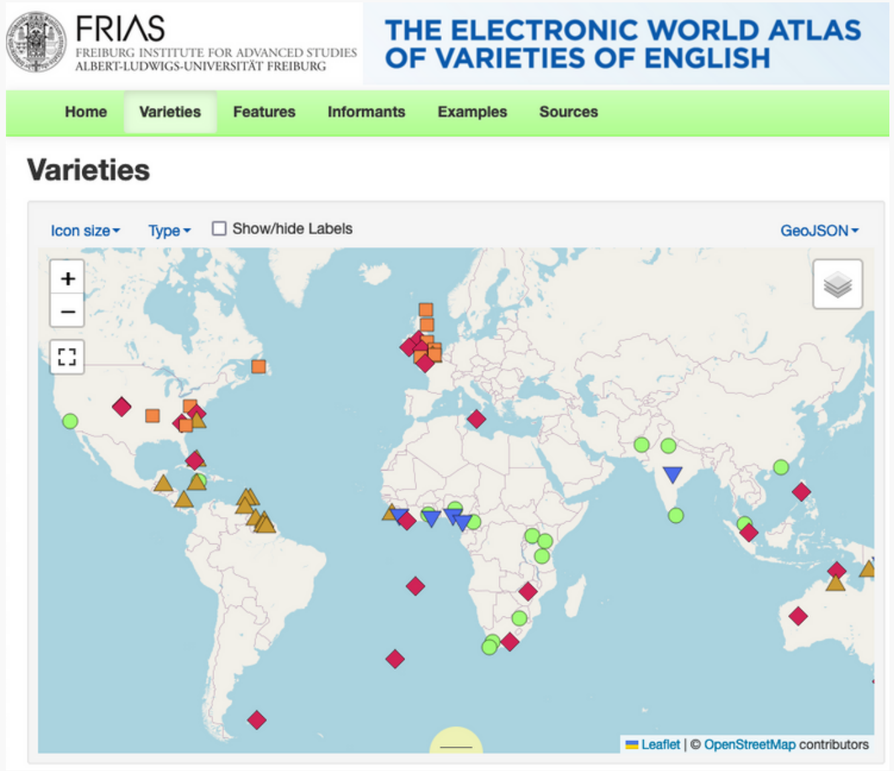


II Overview

- We craft and manually evaluate perturbation rules that transform German sentences into colloquial forms and use them to synthesize test sets in four ToD datasets;
- Our perturbation rules cover 18 phenomena;
- We conduct an experimental evaluation across six different transformers.



Syntactic perturbations – English

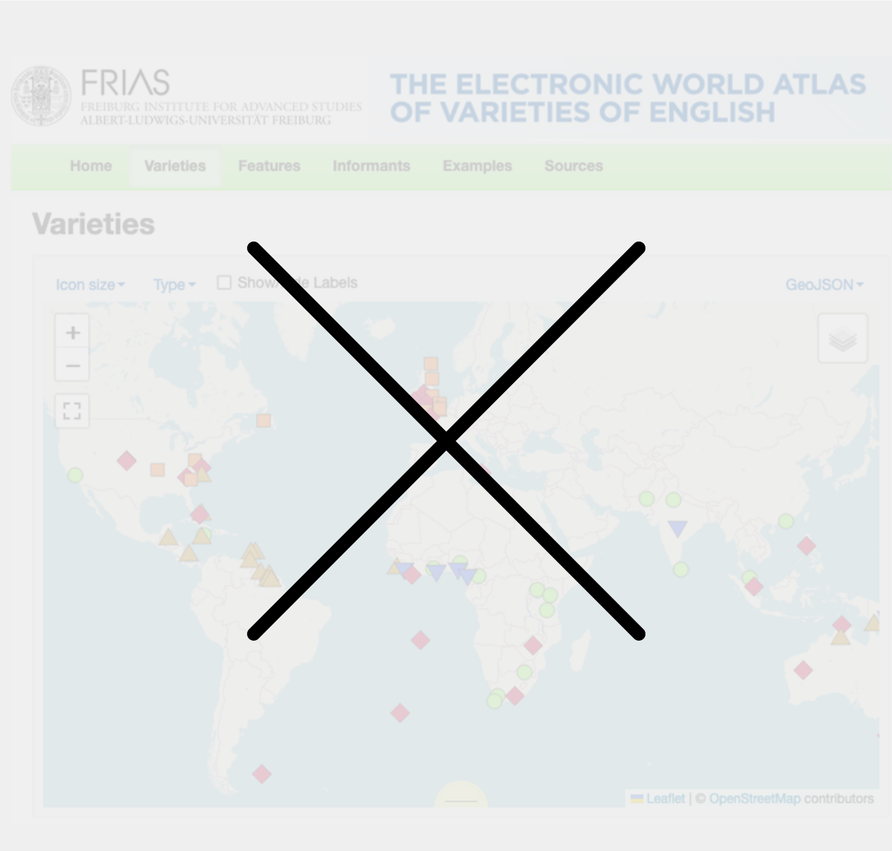


Multi-VALUE: A Framework for Cross-Dialectal English NLP

Caleb Ziems **William Held** **Jingfeng Yang**
Jwala Dhamala **Rahul Gupta** **Diyi Yang**
Stanford University, Georgia Institute of Technology, Amazon
{cziems, diyiy}@stanford.edu, {wheld3}@gatech.edu,
{jddhamal, yjfllpyym, gupra}@amazon.com

→ Rules for perturbing the syntax of Standard American English
to mimic the structure of other varieties

Syntactic perturbations – German



Review >30 works on German dialect syntax
→ 18 rules covering both widespread and local syntactic features

Our perturbations target:

- possessive constructions, determiner structures, comparatives, aspect, negation, personal names, prepositions...

Syntactic perturbations – German

Spiele	**das**	**Lied**	**von**	**Eddie**	**Vinson**
Play.IMP	the	song	by	Eddie	Vinson
Intent: PlayMusic		*Slot: music item*		*Slot: artist*	

Syntactic perturbations – German

	Spiele	**das**	**Lied**	**von**	**Eddie**	**Vinson**
	Play.IMP	the	song	by	Eddie	Vinson
Swap name order	**Spiele**	**das**	**Lied**	**von**	**Vinson**	**Eddie**
	Play.IMP	the	song	by	Vinson	Eddie

Syntactic perturbations – German

	Spiele	**das**	**Lied**	**von**	**Eddie**	**Vinson**
	Play.IMP	the	song	by	Eddie	Vinson
Swap name order	**Spiele**	**das**	**Lied**	**von**	**Vinson**	**Eddie**
	Play.IMP	the	song	by	Vinson	Eddie
Article before personal name	**Spiele**	**das**	**Lied**	**vom**	**Vinson**	**Eddie**
	Play.IMP	the	song	by.the.DAT	Vinson	Eddie

Syntactic perturbations – German

	Spiele	**das**	**Lied**	**von**	**Eddie**	**Vinson**	
	Play.IMP	the	song	by	Eddie	Vinson	
Swap name order	**Spiele**	**das**	**Lied**	**von**	**Vinson**	**Eddie**	
	Play.IMP	the	song	by	Vinson	Eddie	
Article before personal name	**Spiele**	**das**	**Lied**	**vom**	**Vinson**	**Eddie**	
	Play.IMP	the	song	by.the.DAT	Vinson	Eddie	
Imperative with "tun" construction	**Tue**	**das**	**Lied**	**vom**	**Vinson**	**Eddie**	**spielen**
	Do.IMP	the	song	by.the.DAT	Vinson	Eddie	play.INF

IV Research Questions

RQ1: How does the LM performance in intent recognition and slot filling change when applied to synthetic dialectal data?

RQ2: As each perturbation isolates a specific phenomenon, which perturbations have the most significant effect?

RQ3: How do LMs differ in terms of robustness to dialectal perturbations?

Datasets

xSID (van der Goot et al., 2021)

- no DEU test
- general domain
- 16 intent classes, 33 slot types
- 300 / 500 samples in dev/test sets

MultiATIS++ (Xu et al., 2020)

- aviasales
- 18 intent classes, 84 slot types
- 1.2k / 893 samples in dev/test sets

MASSIVE (Bastianelli et al., 2020)

- smart home
- 60 intent classes, 55 slot types
- 2k / 3k samples in dev/test sets

MTOP (Li et al., 2021)

- virtual assistant
- 117 intent classes, 78 slot types
- 1.8k/ 3.5k samples in dev/test sets

Models

- mBERT (Devlin et al., 2019)
  - Bavarian Wikipedia is in pre-training data
- RemBERT
- DistilmBERT (Sanh et al., 2019)
- XLM-R (Conneau et al., 2020)
- mDeBERTa (He et al., 2021)
- mMiniLM (Wang et al., 2020)

Evaluation metrics

Intent recognition accuracy

Slot filling F1

- span and label must match exactly

Drop in performance metrics before and after the perturbations are applied

Attack success rate

- the number of instances that become misclassified after the perturbation is applied

VI SETUP

Train on std ENG
Validate on std ENG
Test on std + dialect DEU

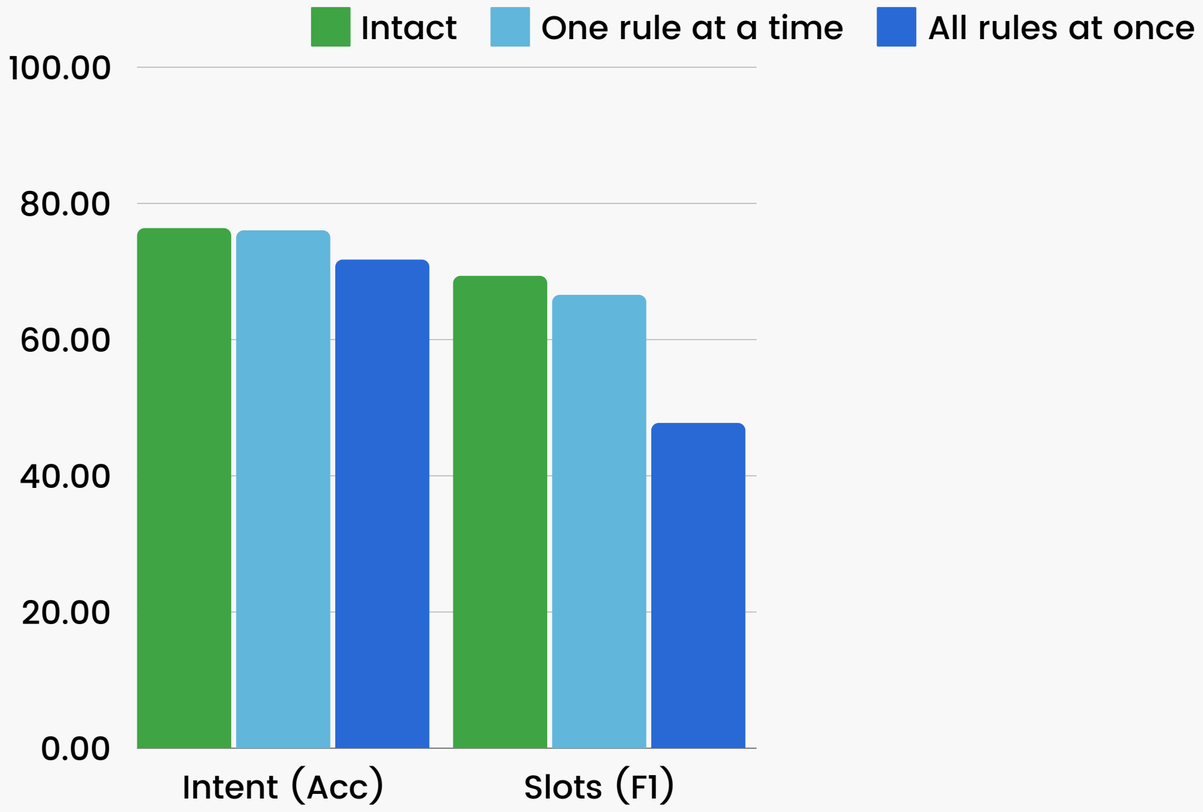
Train on std ENG
Validate on std DEU
Test on std + dialect DEU

Train on std DEU
Validate on std DEU
Test on std + dialect DEU

- joint intent recognition and slot-filling
- train with MaChAmp (van der Goot et al., 2021)
- 5 random seeds
- average all metrics
- use a single GPU

VII Results

RQ1: How does the LM performance in intent recognition and slot filling change when applied to synthetic dialectal data?



ENG train & dev → DEU test

(Sentence-level) intent recognition robust,
(word-level) slot filling brittle

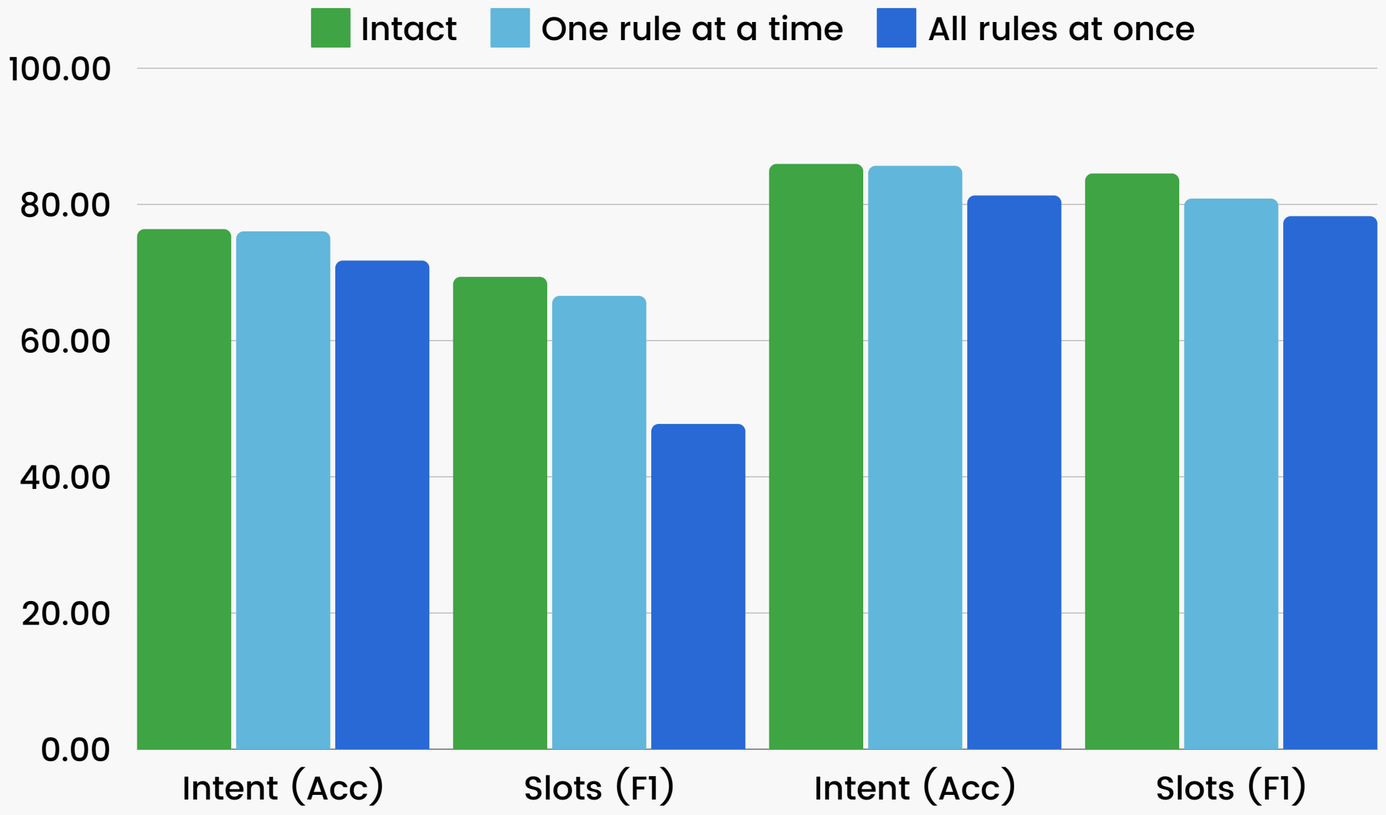
Similar trends for

- ENG train, DEU dev
  → DEU test

(mean scores; details per model & dataset in paper)

VII Results

RQ1: How does the LM performance in intent recognition and slot filling change when applied to synthetic dialectal data?



ENG train & dev → DEU test

DEU train & dev → DEU test

(Sentence-level) intent recognition robust,
(word-level) slot filling brittle

Similar trends for

- ENG train, DEU dev
  → DEU test

(mean scores; details per model & dataset in paper)

VII Results

RQ2: As each perturbation isolates a specific phenomenon, which perturbations have the most significant effect?

ENG train & dev → DEU test

Intent recognition

- Largest (negative) impact on model performance:
  - Swapping first and last name

Slot filling

- Perturbations altering the word order have the greatest impact
- Then changes to noun and verb phrases
- In MultiATIS++ (travel-planning), changes to direction/location prepositions are impactful

Rarely seen dialect phenomena deceive models more effectively.

VII Results

RQ3: How do LMs differ in terms of robustness to dialectal perturbations?

mDeBERTa RemBERT	best performance across the board
XLM-R	sometimes competitive, sometimes clearly worse
mBERT DistilmBERT mMiniLM	worst performance

All models are affected by the perturbations

VII Results

Error analysis

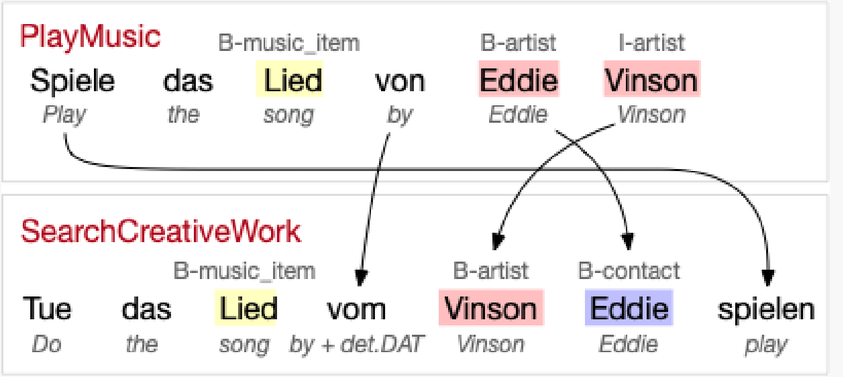
Confused intents

- similar intents (play music, search creative work)
- intents with homonymous associated terms (book a table, rate a book)

→ more pronounced when perturbations are applied

Slot filling issues

- Changed word order → slot boundary issues
- Incorrect slot types, e.g., when prepositions are changed
- Words added (periphrastic verb constructions)
  → extra slot labels assigned



VIII Conclusion

Possible future directions

- Spoken language understanding and modelling phonological phenomena
- Incorporating lexical variation by relying on bilingual lexicons

VIII Conclusion

We encourage the community to take on experiments with various languages and dialects

- fair evaluation approaches, that account for dialects and don't favor standard languages;
- a better understanding of specific requirements of dialect speakers.

github.com/mainlp/dialect-ToD-robustness