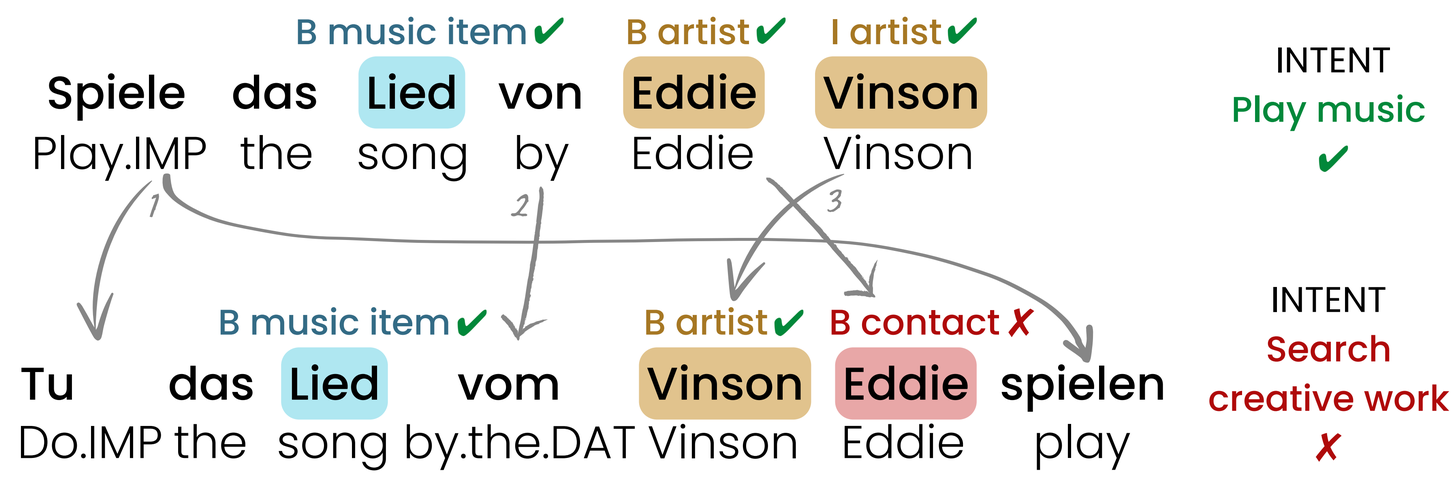
Ekaterina Artemova
Verena Blaschke
Barbara Plank
MaiNLP lab, CIS, LMU Munich

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

Typical set-up: train a model on ENG intent recognition + slot-filling and apply it zero-shot to other languages.
What about lower-resource colloquial varieties with limited test data?



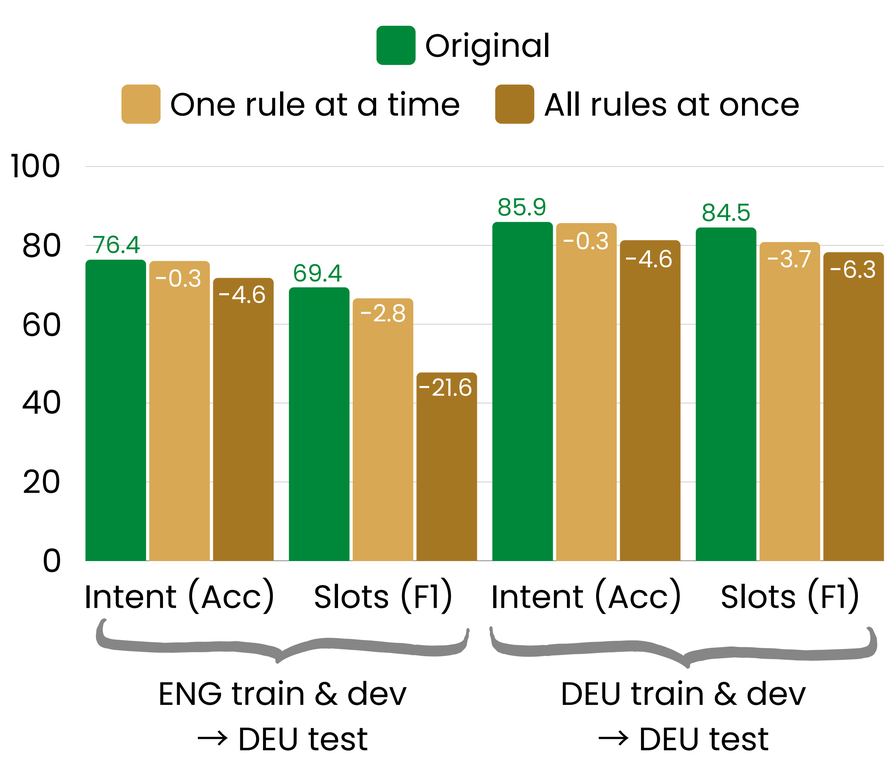
1 Periphrastic imperative with auxiliary *tun* ("do")
2 Article before personal names (here merged with apposition)
3 Swapped order of given and family name

Method

- 18 handcrafted and validated morphosyntactic perturbation rules for German dialects & colloquial varieties (inspired by Ziems et al. 2023) targeting
  - genitive constructions
  - personal names
  - verb clusters
  - prepositions
  - relative sentences
  - & more
- 4 datasets: xSID, MultiATIS++, MASSIVE, MTOP
- 5 multilingual LMs: mBERT, XLM-R, RemBERT, mDeBERTa, DistilmBERT, mMiniLM
- Training: joint intent recognition and slot-filling, 5 random seeds

How does the LMs' performance change when applied to synthetic colloquial data?

- Intent recognition (sentence-level): robust
- Slot-filling (word-level): brittle, especially when multiple perturbations are applied at once



Which perturbations have the most significant effect? What kinds of errors do we get?

Intent recognition
- Biggest impact: swapping first/last name
- Confusion between intents with similar meanings or similar associated terms

Slot filling
- Biggest impact: changes to the word order → issues with slot boundaries
- In MultiATIS++ (travel-planning), changes to direction/location prepositions are impactful → incorrect slot types
- Extra slot labels get assigned when words are added (e.g. in periphrastic verb constructions)

Rarely seen dialect phenomena deceive models more effectively

Which LMs are the most/least robust?

All models are affected!
- Best across the board: mDeBERTa, RemBERT
- Sometimes competitive: XLM-R
- Worst: mBERT, DistilmBERT, mMiniLM

Conclusion

Linguistic variation has an effect on slot/intent detection performance, even when only morphosyntax is concerned!
Future work: what if we also change lexis & pronunciation?

For additional details, check out the paper & code!

Munich Center for Machine Learning
European Research Council
Established by the European Commission

This research is supported by the ERC Consolidator Grant DIALECT 101043235 and the Munich Center for Machine Learning.