Improving dialectal slot and intent detection with auxiliary tasks

A multi-dialectal Bavarian case study

Xaver Maria Krückl*, **Verena Blaschke*** & Barbara Plank
LMU Munich



VarDial
January 19, 2025

Intro: Slot and intent detection

Remind me to call Stephanie on Tuesday → *set reminder*

reminder todo *datetime*

- Almost(?) solved for English datasets
- What if we speak another language?

Erinner mich Stephanie am Dienstag anzurufen

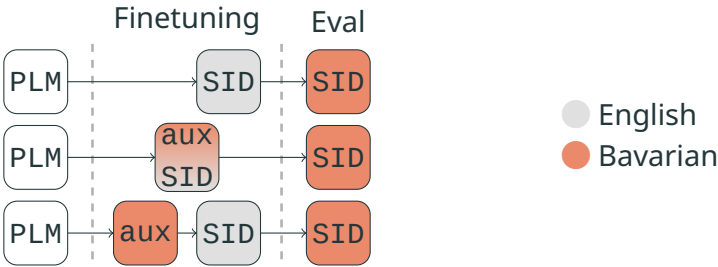
- German: no(t much) training data – but MT, multilingual LMs

Erinner mi d Stephanie in Dienschtig unzuriafn

- Bavarian dialects: no training data

Intro: Auxiliary tasks

No SID training data … but other Bavarian datasets!



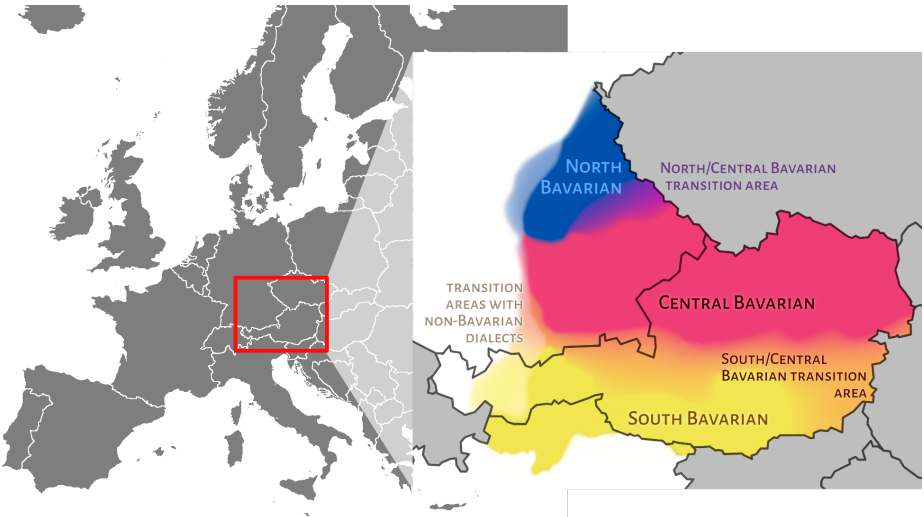
Contributions

1. New Bavarian SID dataset
2. Effect of dialectal auxiliary tasks
3. Robustness of results

Overview

1. **New Bavarian SID dataset**
2. Effect of dialectal auxiliary tasks
3. Robustness of results

Background: Bavarian dialects

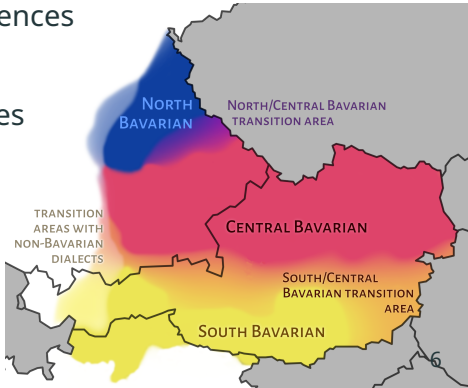


Background: Bavarian dialects

Closely related to German, but

- Phonological and phonetic differences
- Lexical differences
- Some morphosyntactic differences
- No orthography

→ All of this is reflected in our data!



(Bavarian) SID data

xSID

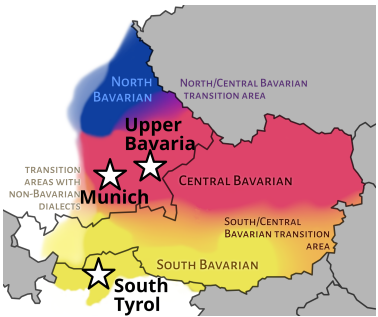
- 16 intents, 33 slot types
- English training data
- Multilingual evaluation data, incl. 2 Bavarian dialects

van der Goot+ (NAACL 2021),
Aepli+ (VarDial 2023),
Winkler+ (LREC-COLING 2024)

Bavarian SID data

New $\rightarrow$ to be included in xSID

- Munich Bavarian (native speaker, 20s)
- Comparison of dialectal/orthographic/translation styles



Munich streich olle wecka

U. Bav. Lösch olle Wegga

S. Tyrol tua olle Wecker weck

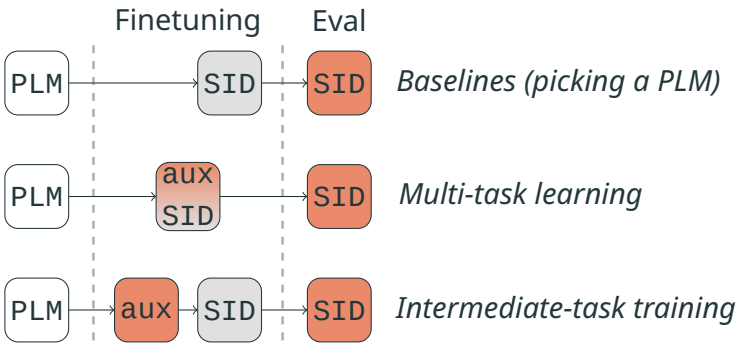
"Delete all alarms"

Overview

1. New Bavarian SID dataset
2. **Effect of dialectal auxiliary tasks**
3. Robustness of results

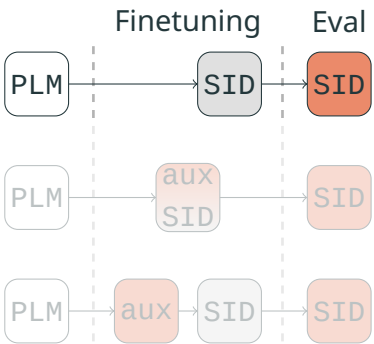
Set-ups

- Train on English ⚪ SID data (+ Bavarian auxiliary tasks)
- Test on the 3 Bavarian dialects 🟠 (avg)



Set-ups

- Train on English ⚪ SID data (+ Bavarian auxiliary tasks)
- Test on the 3 Bavarian dialects 🟠 (avg)

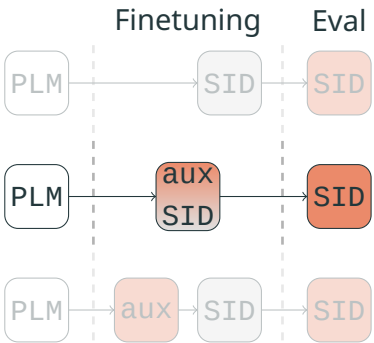


Baselines: transfer from English (Base-size) PLMs

- 1× German (GBERT)
- 3× multilingual (mBERT, XLM-R, mDeBERTa)

Set-ups

- Train on English ⚪ SID data (+ Bavarian auxiliary tasks)
- Test on the 3 Bavarian dialects 🟠 (avg)

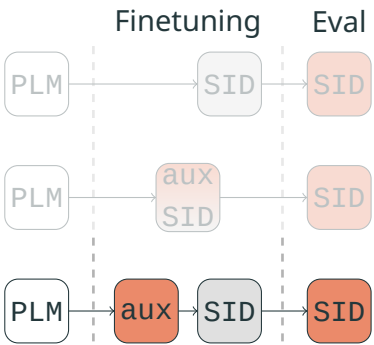


Multi-task learning
[task] × SID

Simultaneously learn from
the English SID data
& 1+ Bavarian task

Set-ups

- Train on English ⚪ SID data (+ Bavarian auxiliary tasks)
- Test on the 3 Bavarian dialects 🟠 (avg)

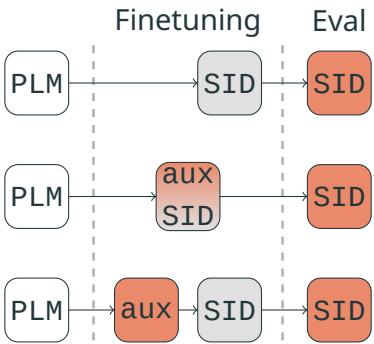


Intermediate-task training
[task] → SID

First, train on Bavarian NLP tasks
Then, train on English SID data

Set-ups

- Train on English ⚪ SID data (+ Bavarian auxiliary tasks)
- Test on the 3 Bavarian dialects 🟠 (avg)

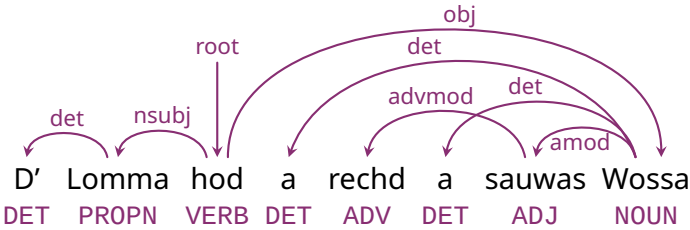


Related work

- Mostly monolingual
- Mostly either MTL or ITT
- No dialects/non-std languages

Auxiliary tasks

Syntax



“The (river) Lammer has fairly clean water”

(Joint) dependency parsing + POS tagging

- ∼900 sentences
- Many Bavarian dialects
- Linguistic structure useful for identifying slots?

MaiBaam (Blaschke+, LREC-COLING 2024)

Auxiliary tasks

Named entity recognition

person *organization*

Da Rudoif hod 1365 de Universität Wean grindt

“Rudolf founded the University of Vienna in 1365”

- 9k sentences
- Multiple Bavarian dialects
- Similar to slot filling?

BarNER (Peng+, LREC-COLING 2024)

Auxiliary tasks

Masked language modelling

Waun i du wa, tarat i `[MASK]` frogn

↓

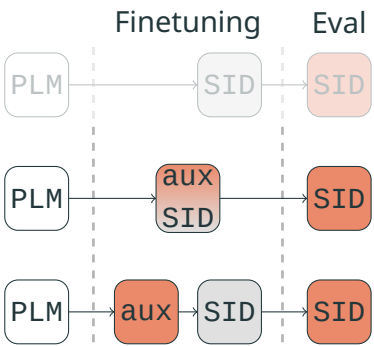
’n

“If I were you, I’d ask him”

- ⚠ Small dataset (1.4k sentences)
- Multiple Bavarian dialects
- Common pre-training objective

Wikipedia data via Artemova/Plank (NoDaLiDa 2023)

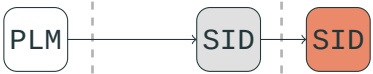
Auxiliary tasks

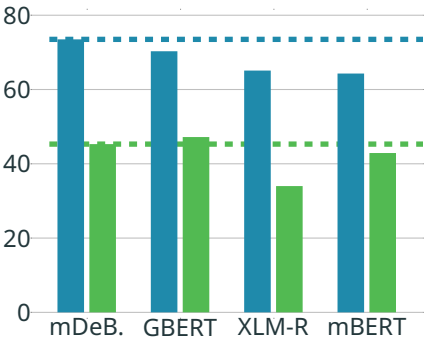


- Each task once for MTL (Aux$\times$SID)
- Once for ITT (Aux$\rightarrow$SID)
- Some task combinations (not exhaustive)

3 random seeds

Results: Baselines





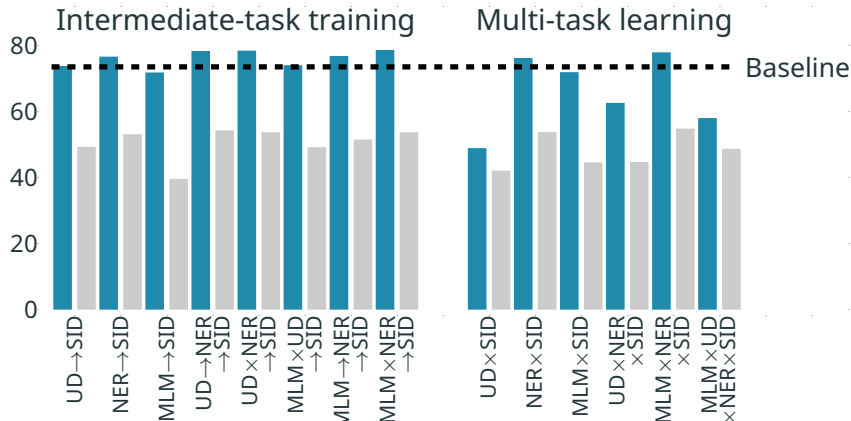
- Best: Monolingual + mDeBERTa
- Continue w/ mDeBERTa (multilingual)
  - 73.5 % intent acc.
  - 45.3 % slot F1

Results: General trends

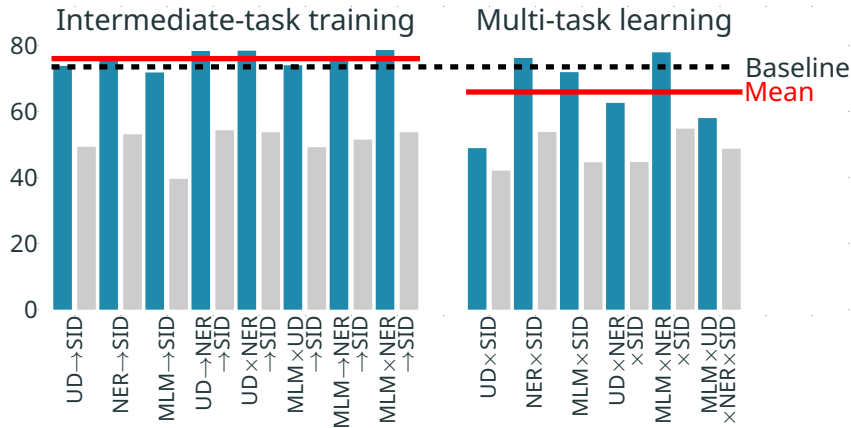
Set-ups with auxiliary tasks

- More improvements for slots than for intents
  - Auxiliary tasks: token level
  - (Almost) all set-ups that improve slot filling also improve intent classification (+ vice versa)
- Results depend both on the set-up (joint/sequential) and on the exact tasks used

Results: Joint vs. sequential learning – Intents

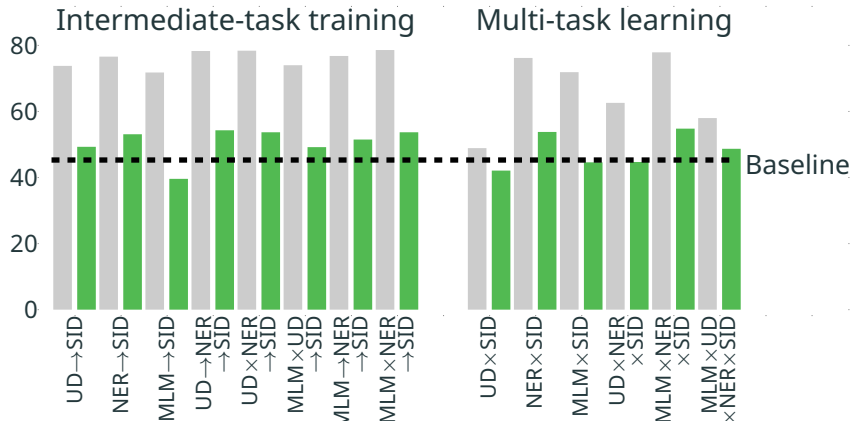


Results: Joint vs. sequential learning – Intents

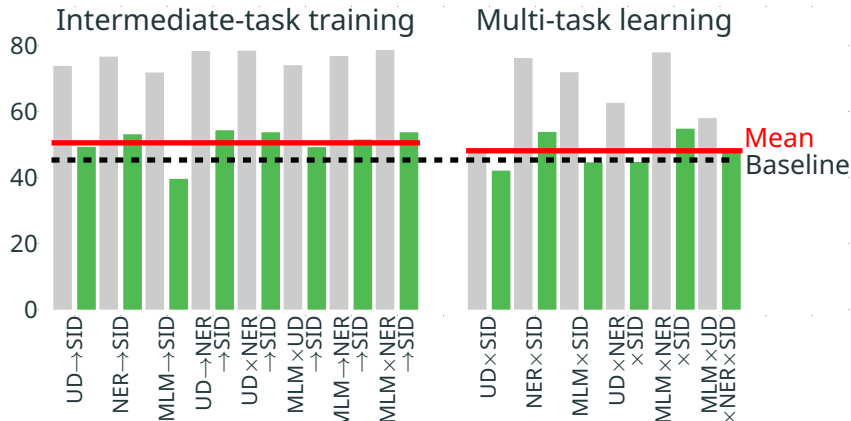


- Most aux. tasks are also learned better on their own (but MTL with NER improved POS/parsing scores)

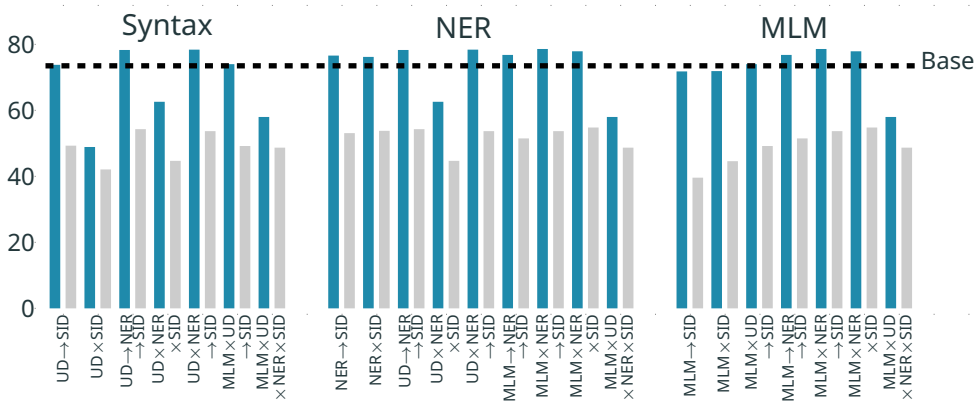
Results: Joint vs. sequential learning – Slots



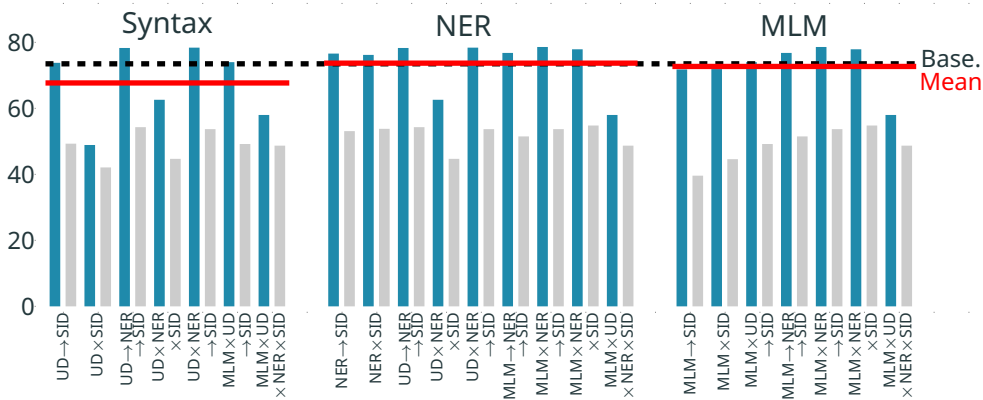
Results: Joint vs. sequential learning – Slots



Results: Auxiliary task choice – Intents

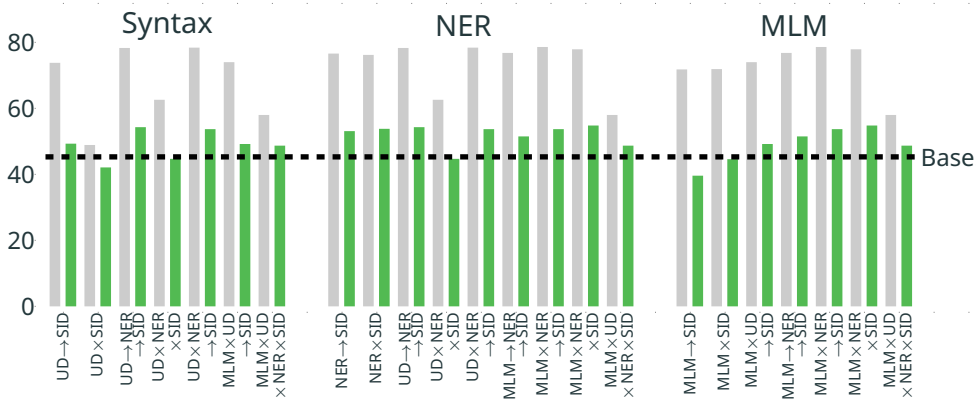


Results: Auxiliary task choice – Intents

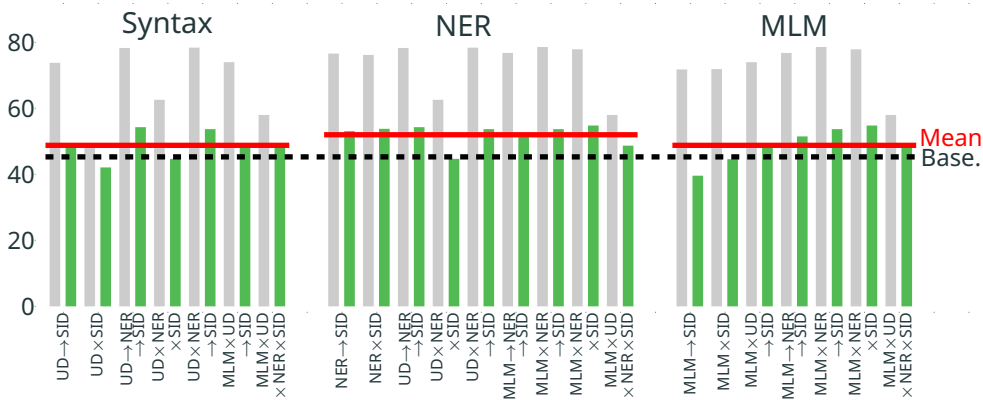


- Syntax: only harmful in joint multi-task learning set-ups!

Results: Auxiliary task choice – Slots



Results: Auxiliary task choice – Slots



Results: One task or multiple aux. tasks?

Combined *can* be more helpful than individual

	Intents	Slots
Baseline (SID)	73.5	45.3
UD $\rightarrow$ SID	+0.3	+3.9
NER $\rightarrow$ SID	+3.0	+7.8
UD $\times$ NER $\rightarrow$ SID	+4.8	+8.4
UD $\rightarrow$ NER $\rightarrow$ SID	+4.8	+9.0

Best models involve multiple aux. tasks

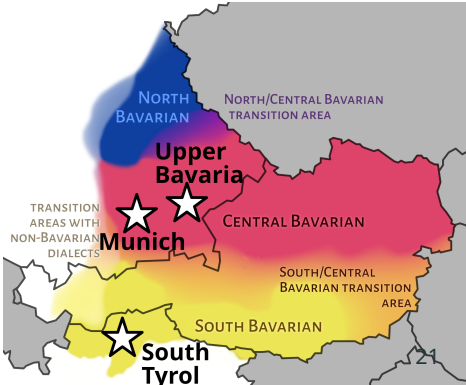
- MLM $\times$ NER $\rightarrow$ SID – intents: 78.6%, slots: 53.7%
- UD $\times$ NER $\rightarrow$ SID
- UD $\rightarrow$ NER $\rightarrow$ SID

Overview

1. New Bavarian SID dataset
2. Effect of dialectal auxiliary tasks
3. **Robustness of results**

Results: Differences across dialects

- Best results for Upper Bavarian, then South Tyrolean, then Munich
- Aux. data: *Very* stable trends across Bavarian dialects
- Also similar impacts on German test data (but less strong)



Robustness across datasets

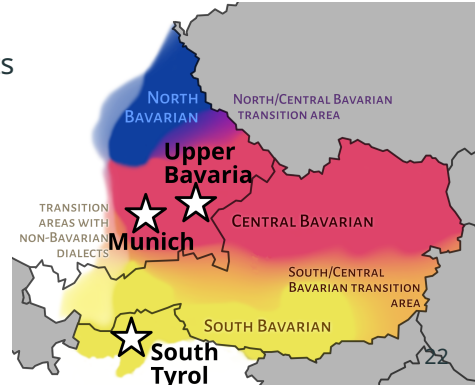
Additional test sets for Upper Bavaria

- Naturalistic data
- Translation of MASSIVE

Subset of models

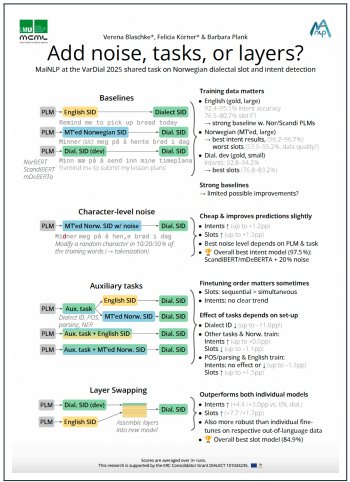
→ Substantial drops across datasets
→ Aux. task trends: not the same, but similar
→ Gains: up to 7 pp. for intents up to 10 pp. for slots

NaLiBaSID (Winkler+, LREC-COLING 2024), MASSIVE (FitzGerald+, ACL 2023)



Beyond this study

Results appear to depend somewhat on the actual datasets, languages, baseline performances…

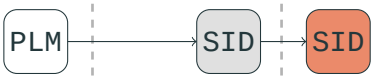
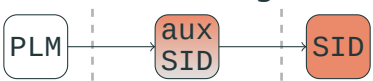
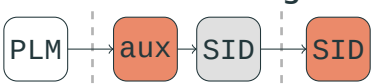


Conclusion

- Additional tasks can improve/decrease the performance
- Greatest gains...
  - For slot filling
  - With intermediate-task training (sequential)
  - With NER
- New Bavarian dev/test data translation

Thank you!

Appendix: Results – Joint vs. sequential learning

	Intents	Slots
Baseline 	73.5% acc.	45.3% F1
Multi-task learning 	mean –7.6 pp. (–24.6 – +4.3)	mean +2.8 pp. (–3.2 – +9.5)
Interm.-task training 	mean +2.5 pp. (–1.6 – +5.1)	mean +5.2 pp. (–5.8 – +9.0)

Appendix: Results – Auxiliary task choice

	Intents	Slots
Baseline	73.5 % acc.	45.3% F1
Syntax	mean –5.8 pp. (–24.6 – +4.8)	mean +3.5 pp. (–3.2 – +9.0)
NER	mean +0.2 pp. (–15.6 – +5.1)	mean +6.7 pp. (–0.6 – +9.5)
MLM	mean –0.8 pp. (–15.6 – +5.1)	mean +3.5 pp. (–5.8 – +9.5)

Appendix: Detailed results

	Intents						Slots					
	Avg	ITT	MTL	UD	NER	MLM	Avg	ITT	MTL	UD	NER	MLM
SID (mDeBERTa)	73.5						45.3					
UD→SID	73.8	+0.3		+0.3			49.3	+3.9		+3.9		
UD×SID	48.9		–24.6	–24.6			42.1		–3.2	–3.2		
NER→SID	76.5	+3.0			+3.0		53.1	+7.8			+7.8	
NER×SID	76.2		+2.7		+2.7		53.8		+8.4		+8.4	
MLM→SID	71.8	–1.8				–1.8	39.6	–5.8				–5.8
MLM×SID	71.9		–1.6			–1.6	44.6		–0.7			–0.7
UD→NER→SID	78.3	+4.8		+4.8	+4.8		54.3	+9.0		+9.0	+9.0	
UD×NER→SID	78.4	+4.8		+4.8	+4.8		53.7	+8.4		+8.4	+8.4	
UD×NER×SID	62.6		–10.9	–10.9	–10.9		44.7		–0.6	–0.6	–0.6	
MLM×UD→SID	73.9	+0.4		+0.4		+0.4	49.2	+3.8		+3.8		+3.8
MLM→NER→SID	76.8	+3.3			+3.3	+3.3	51.5	+6.2			+6.2	+6.2
MLM×NER→SID	78.6	+5.1			+5.1	+5.1	53.7	+8.4			+8.4	+8.4
MLM×NER×SID	77.9		+4.3		+4.3	+4.3	54.8		+9.5		+9.5	+9.5
MLM×UD×NER×SID	58.0		–15.6	–15.6	–15.6	–15.6	48.7		+3.3	+3.3	+3.3	+3.3
Mean		+2.5	–7.6	–5.8	+0.2	–0.8		+5.2	+2.8	+3.5	+6.7	+3.5

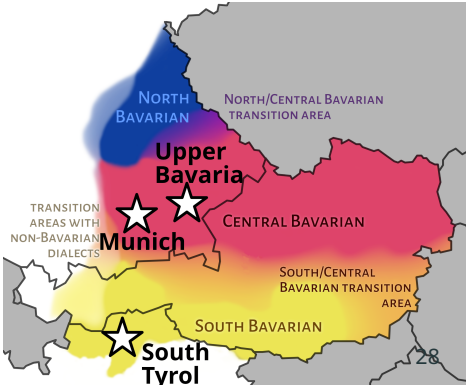
Robustness across datasets

Additional test sets for Upper Bavaria

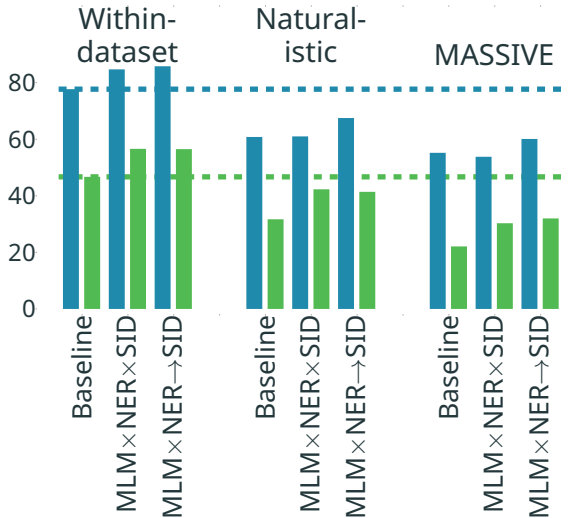
- Naturalistic data
- Translation of MASSIVE

Model comparisons

- Baseline
- Overall best model (MLM×NER→SID)
- Its fully MTL counterpart (MLM×NER×SID)



Results: Robustness across datasets



- Substantial drops across datasets
- Aux. task trends not exactly the same, but similar