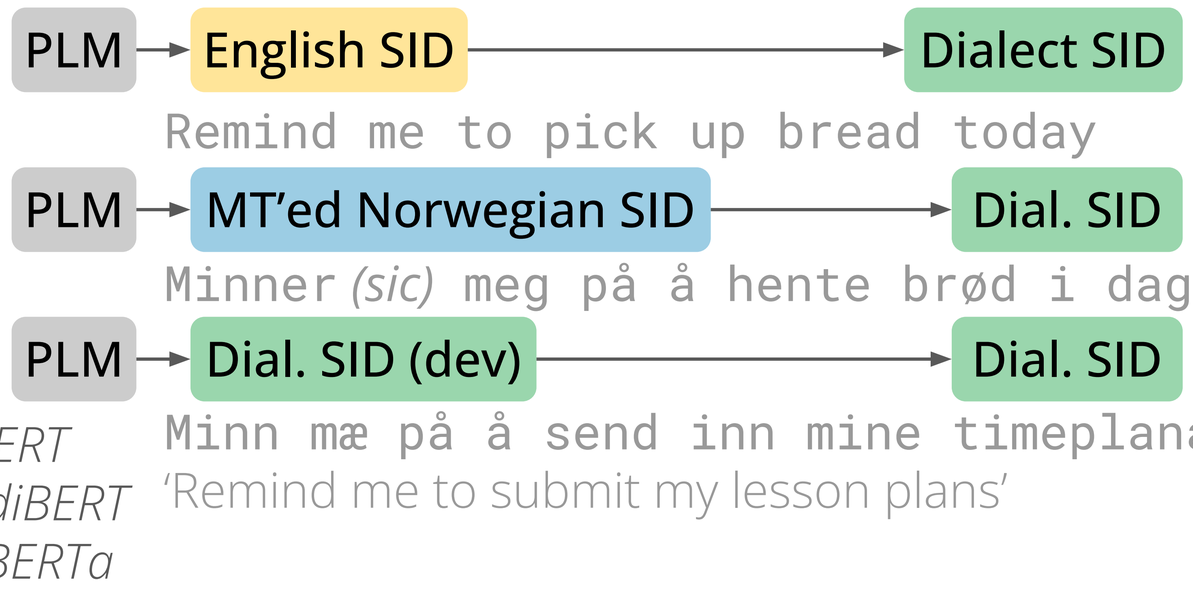
Verena Blaschke*, Felicia Körner* & Barbara Plank

# Add noise, tasks, or layers?

MaiNLP at the VarDial 2025 shared task on Norwegian dialectal slot and intent detection

## Baselines



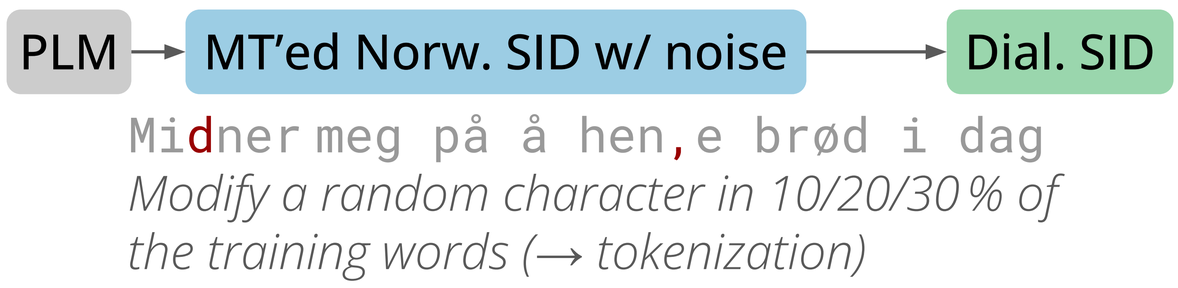

### Training data matters

- English (gold, large)
  92.4–95.1% intent accuracy
  76.5–80.7% slot F1
  → strong baseline w. Nor/Scandi PLMs
- Norwegian (MT'ed, large)
  → best intent results, (96.2–96.7%)
  worst slots (53.9–55.2%, data quality?)
- Dial. dev (gold, small)
  intents: 92.8–94.2%
  → best slots (76.8–83.2%)

### Strong baselines

→ limited possible improvements?

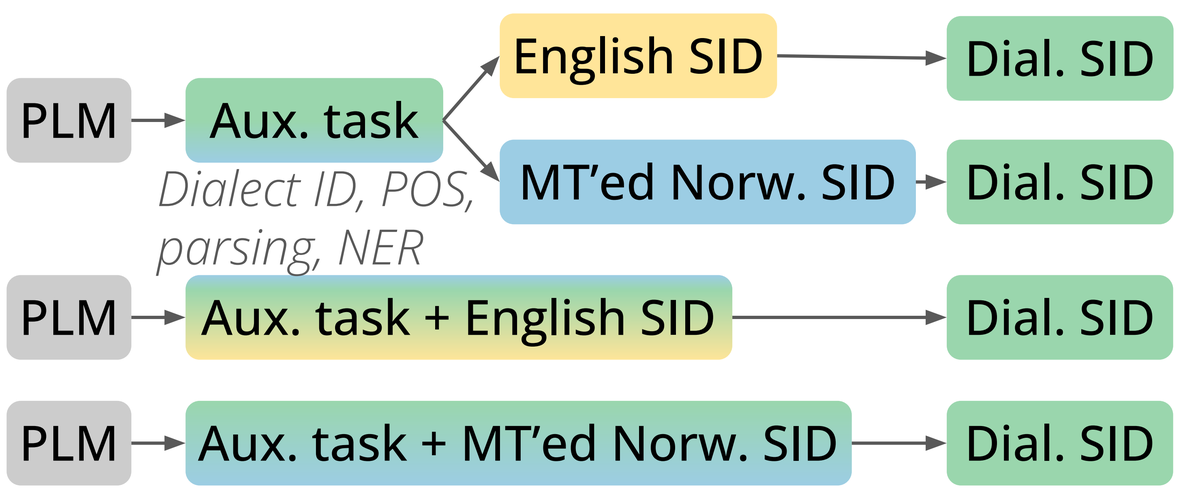
## Character-level noise




### Cheap & improves predictions slightly

- Intents ↑ (up to +1.2pp)
- Slots ↑ (up to +1.3pp)
- Best noise level depends on PLM & task
- 🏆 Overall best intent model (97.5%):
  ScandiBERT/mDeBERTA + 20% noise

## Auxiliary tasks



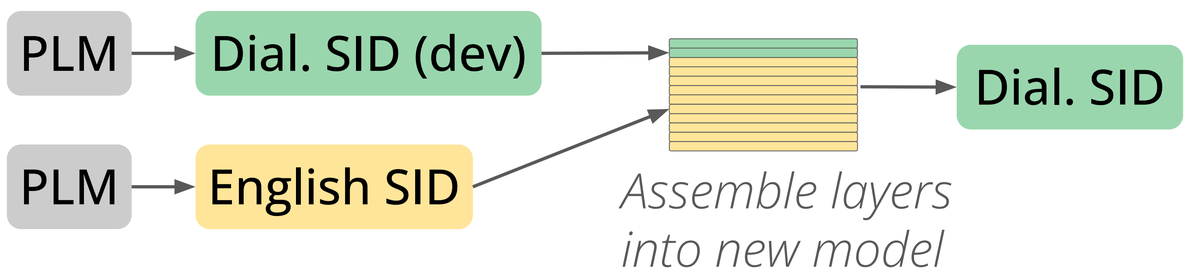

### Finetuning order matters sometimes

- Slots: sequential > simultaneous
- Intents: no clear trend

### Effect of tasks depends on set-up

- Dialect ID ↓ (up to –11.0pp!)
- Other tasks & Norw. train:
  Intents ↑ (up to +0.6pp)
  Slots ↓ (up to –1.1pp)
- POS/parsing & English train:
  Intents: no effect or ↓ (up to –1.3pp)
  Slots ↑ (up to +1.5pp)

## Layer Swapping




### Outperforms both individual models

- Intents ↑ (+4.4 /+3.0pp vs. EN, dial.)
- Slots ↑ (+7.7 /+1.7pp)
- Also more robust than individual fine-tunes on respective out-of-language data
- 🏆 Overall best slot model (84.9%)

Scores are averaged over 3+ runs.

This research is supported by the ERC Consolidator Grant DIALECT 101043235.