Add Noise, Tasks, or Layers? MaiNLP at the VarDial 2025 Shared Task on Norwegian Dialectal Slot and Intent Detection

Verena Blaschke* **Felicia Körner*** **Barbara Plank**
MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
Munich Center for Machine Learning (MCML), Munich, Germany
{verena.blaschke, f.koerner, b.plank}@lmu.de

Abstract

Slot and intent detection (SID) is a classic natural language understanding task. Despite this, research has only more recently begun focusing on SID for dialectal and colloquial varieties. Many approaches for low-resource scenarios have not yet been applied to dialectal SID data, or compared to each other on the same datasets. We participate in the VarDial 2025 shared task on slot and intent detection in Norwegian varieties, and compare multiple set-ups: varying the training data (English, Norwegian, or dialectal Norwegian), injecting character-level noise, training on auxiliary tasks, and applying Layer Swapping, a technique in which layers of models fine-tuned on different datasets are assembled into a model. We find noise injection to be beneficial while the effects of auxiliary tasks are mixed. Though some experimentation was required to successfully assemble a model from layers, it worked surprisingly well; a combination of models trained on English and small amounts of dialectal data produced the most robust slot predictions. Our best models achieve 97.6% intent accuracy and 85.6% slot F_1 in the shared task.

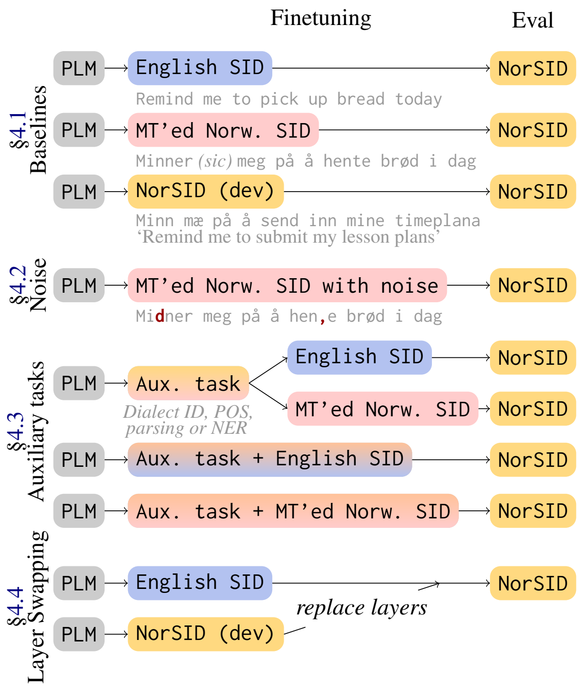

Figure 1: **Overview of our approaches:** pre-trained language models (PLMs) fine-tuned on English, machine-translated Norwegian data or the dialectal development set; noise injection into the Norwegian data; training on auxiliary tasks in addition to SID data (sequentially or jointly); assembling layers of models fine-tuned on different datasets.

1 Introduction

Slot and intent detection (SID) is a classic natural language understanding (NLU) task. Research today has mainly focused on standard languages with many speakers (e.g., Schuster et al., 2019; Xu et al., 2020; Li et al., 2021; FitzGerald et al., 2023). However, even when performance on a related standard language is high, SID for non-standard varieties can be challenging. This can be due to spelling variation (Srivastava and Chiang, 2023b) and syntactic differences that complicate cross-lingual slot filling (Artemova et al., 2024). Furthermore, the lack of task data in the relevant language varieties complicates the adaptation of SID models to under-resourced varieties.

In this paper, we report on the results of our participation in the VarDial 2025 shared task on slot and intent detection in Norwegian standard and dialect varieties (NorSID; Scherrer et al., 2025). We compare multiple strategies for improving the performance of SID systems (Figure 1):

1. Fine-tuning models on large amounts of gold-standard English or silver-standard Norwegian data, or smaller amounts of gold-standard Norwegian dialect data (§4.1);

*Equal contribution.

Proceedings of the Twelfth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 182–199
January 19, 2025.

2. Adding noise to the Norwegian training data to make models more robust to spelling variation (§4.2);

3. Additionally training on auxiliary NLP tasks in Norwegian (§4.3);

4. Assembling layers of models fine-tuned on different tasks or languages into a new model to combine their capabilities (§4.4).

We share our code at `https://github.com/mainlp/NorSID`.

2 Related Work

In the past few years, research on SID for dialects and non-standard languages has gained popularity. The multilingual SID dataset xSID (van der Goot et al., 2021a; Aepli et al., 2023; Winkler et al., 2024) contains evaluation data in over a dozen languages, including non-standard varieties like Neapolitan, and German dialects spoken in Switzerland, South Tyrol, and Bavaria. It has recently been extended with translations into Norwegian dialects (Mæhlum and Scherrer, 2024), which are the focus of this shared task. We provide more details in §3.

Using xSID, van der Goot et al. (2021a) investigate multi-task learning with auxiliary tasks in the target language (or a closely related standard language). Similarly, Krückl et al. (2025) include auxiliary tasks in multi-task learning and intermediate-task training set-ups for dialectal SID. Both studies find that the effects depend on both the auxiliary task(s) and the target task. We include auxiliary tasks in one of our experiments (§4.3).

Two previous shared tasks have focused on dialectal SID (Aepli et al., 2023; Malaysha et al., 2024). Useful approaches by the participants were to train on SID data in multiple languages (Kwon et al., 2023), injecting character-level noise into the training data (Srivastava and Chiang, 2023b; we use a similar method in §4.2), and ensembling models trained on dialectal translations of the training set (Ramadan et al., 2024; Elkordi et al., 2024; Fares and Touileb, 2024).

Outside the context of a shared task, Abboud and Oz (2024) also focus on generating synthetic dialectal training data. Lastly, Muñoz-Ortiz et al. (2025) find that visual input representations are more robust than subword token embeddings when transferring German intent classification models to related dialects.

Label type / data subset	Train	Dev	Test
Intents	18	15	15
Slot types	40	33	34
English	43k	(not used)	
Bokmål	(MT) 43k	1×300	1×500
North Norwegian	—	2×300	2×500
Trønder Norwegian	—	3×300	3×500
West Norwegian	—	5×300	5×500
Total (evaluation)	—	3 300	5 500
Training on dev	2 970	330	5 500

Table 1: **Distribution of labels and languages/dialects in the data.** While 15 intent types occur in both the development and test splits, only 14 of them overlap.

Numerous other methods for improving NLP performance in low-resource settings exist (Hedderich et al., 2021), many of which have not yet been applied to dialectal or cross-lingual SID. One recently proposed approach is assembling layers of models trained on different tasks or languages into a new model (Bandarkar et al., 2024), which we explore in §4.4.

3 Data

We use the xSID 0.6 dataset (van der Goot et al., 2021a) and its Norwegian extension NoMusic (Mæhlum and Scherrer, 2024). xSID combines re-annotated versions of two SID datasets (Coucke et al., 2018; Schuster et al., 2019). It includes 43k English training sentences, as well as smaller development and test datasets that have been translated into other languages. The shared task also includes an automatic translation of the training set into Bokmål (Scherrer et al., 2025). For these sentence-level translations, the intent labels remained unchanged, while the slot annotations were automatically projected during the translation.

NoMusic provides translations of xSID's development and test utterances into the Norwegian Bokmål orthography and Norwegian dialects from three of the four major dialect groups (Table 1). The dialect groups have two to five different translations each. In some of our experiments, we train on the development set, which we split into a training and new development set according to a 90:10 ratio.

Slots are annotated in the `BIO` scheme. Intent classification is measured with accuracy, and slot filling with strict span F_1.

One of our approaches uses datasets for auxiliary tasks; these are described in §4.3. We also use additional datasets for Layer Swapping experiments, described in §4.4.

4 Methodology

We construct several baselines that differ in their training data and pretrained language model (PLM) choices (§4.1). We subsequently build on (some of) these baselines to examine the effects of different recent approaches to improving performance on low-resource language data. We submitted three systems each for intent classification and for slot filling; here we also discuss models we tested but did not submit to the shared task. When selecting systems for submission, we considered their performance on the development set while also aiming for a diverse set of systems.

We use MaChAmp (van der Goot et al., 2021b; v.0.4.2, commit `052044a`) with default hyperparameters to fine-tune the PLMs to simultaneously predict slot and intent labels. The slot predictions are decoded via a conditional random field (CRF; Lafferty et al., 2001). Each model is fine-tuned for 20 epochs, and the best epoch is chosen based on performance on the development set. For Layer Swapping, we build on an implementation of JointBERT (§4.4).

4.1 Baselines

We fine-tune three PLMs as baselines: *1)* the monolingual Norwegian NorBERT v3 (Samuel et al., 2023);[1] *2)* ScandiBERT (Snæbjarnarson et al., 2023),[2] which was pretrained on data in Norwegian, Danish, Swedish, Icelandic, and Faroese; and *3)* mDeBERTa v3 (He et al., 2021, 2023),[3] which was pretrained on 100 languages, including Norwegian (Conneau et al., 2020), and has performed well on dialectal SID data (Artemova et al., 2024; Krückl et al., 2025).

We fine-tune each PLM three times: once on xSID's English training data, once on the machine-translated Norwegian version, and once on NoMusic's development data.

Shared task submission We include the mDeBERTa model trained on the development data in our submissions (slots and intents).[4]

[1] `ltg/norbert3-base` (Apache 2.0)
[2] `vesteinn/ScandiBERT` (AGPL 3.0)
[3] `microsoft/mdeberta-v3-base` (MIT)
[4] `mainlp_{slots,intents}1_mdeberta_siddial_8446`

4.2 Character-Level Noise

Aepli and Sennrich (2022) introduced a simple method for improving transfer from a language to a closely related variety by inserting character-level noise into the training data. Training on the noised data can make a model more robust to spelling variation that results in subword tokenization differences. This method has shown to be beneficial in several studies of transfer to closely related languages and dialects (Aepli and Sennrich, 2022; Srivastava and Chiang, 2023a,b; Brahma et al., 2023; Blaschke et al., 2023, 2024).

We use the machine-translated Norwegian data and randomly select a given percentage of the alphabetic[5] words in a sentence. For each of the selected words, we pick a random position within that word, and delete a character and/or insert one of the alphabetic characters that appear in the Norwegian development set. We implement this once for each of the three PLMs, and compare selecting 10, 20, and 30% of the words.

Shared task submission We submit the mDeBERTa model trained on data with 20% noised words as an intent classification model.[6]

4.3 Auxiliary Tasks

In another set of experiments, we include auxiliary tasks to potentially teach the model tasks related to slot filling and/or relevant information about Norwegian (or Norwegian dialects). Previous studies on training SID models on auxiliary tasks have found that these tasks have different effects on intent detection and slot filling (van der Goot et al., 2021a; Krückl et al., 2025).

Since we are especially interested in whether training on Norwegian auxiliary data can add useful language information to the cross-lingually evaluated English SID model, we use ScandiBERT due to its strong baseline performance when trained on the English SID data. For comparison, we repeat the experiments with the machine-translated Norwegian SID data.

In each of our auxiliary task experiments, we add one additional task to SID. The model parameters are shared across tasks, except for the task-specific decoders. We compare two set-ups: **joint multi-task learning** (where the model is simultaneously learning SID and the other task, cf. Ruder, 2017),

[5] We ignore punctuation and other symbols as well as numbers written as digits.
[6] `mainlp_intents3_mdeberta_sidnor20_5678`

and **intermediate-task training** (where the model is first trained on the auxiliary task, and afterwards on the SID data, cf. Pruksachatkun et al., 2020).

Prior work suggests that choosing auxiliary tasks that are similar to the target task is beneficial for both multi-task learning (Schröder and Biemann, 2020) and intermediate-task training (Poth et al., 2021; Padmakumar et al., 2022). Training on target-language tasks in cross-lingual set-ups has yielded mixed results (van der Goot et al., 2021a; Montariol et al., 2022; Krückl et al., 2025). We include the following auxiliary tasks, which are either in the target dialects or similar to slot filling:

Dialect identification We use NoMusic's development data for dialect classification (with the same 90:10 split as in §4.1) and classify instances on the dialect group level (North Norwegian, Trønder, West Norwegian, or Bokmål).

Part-of-speech tagging and dependency parsing To potentially teach the models about Norwegian sentence structure, we use the part-of-speech (POS) and syntactic dependency annotations of the UD Nynorsk LIA (Øvrelid et al., 2018) treebank. The dataset contains transcriptions of dialectological interviews.[7] We use LIA's phonetic transcriptions and adjust the spelling to be somewhat more natural (Appendix §A). We only include transcribed dialectal material (i.e., exclude utterances by interviewers), leaving 2.3k training and 622 development sentences. We treat POS tagging and dependency parsing as two separate auxiliary tasks.

We note that some of the treebank's dependency annotations violate the Universal Dependencies (de Marneffe et al., 2021) standards and the treebank has been retired from official releases. Nevertheless, we believe that it contains valuable information about Norwegian sentence structure.

Named entity recognition (NER) NER has been useful in other multi-task SID work (Krückl et al., 2025), and gold-standard named entity information has been found to boost slot-filling performance (Yao et al., 2013). As no dialectal NER datasets are available, we use the NorNE dataset (Jørgensen et al., 2020) with a reduced label set (person, organization, location, product, event, derived words).[8] The dataset contains 29.9k training and 4.3k development set sentences (slightly more than half are in Bokmål, and the rest in the other written standard, Nynorsk).

Shared task submission We submit the model first trained on the dependency data and subsequently on xSID's English data for slot filling.[9]

4.4 Layer Swapping

Layer Swapping was recently proposed as a method for cross-lingual transfer (Bandarkar et al., 2024). The authors fine-tune a task expert on English instruction data, and a language expert on general-purpose data in the target language. They replace the top and bottom layers of the task expert with the corresponding layers of the language expert, producing a model capable of performing the task in the target language. We adapt this method – originally applied to LLAMA 3.1 8B (Grattafiori et al., 2024), a 32-layer decoder model – to a 12-layer encoder model.

Experts We use mDeBERTa (He et al., 2023), as it is the strongest baseline when fine-tuned on the NoMusic training data. We replace layers of an **EnSID expert** with layers from a **Norwegian expert**, and consider different options for the latter.

To produce the EnSID expert, we jointly fine-tune on the English xSID training data for both slot filling and intent classification. We build on a JointBERT implementation (closely following Chen et al., 2019),[10] using default hyperparameters, which do not include a CRF and specify 10 epochs. The best checkpoint is chosen based on performance on the NoMusic development set.

We consider four options for the Norwegian expert: the NoMusic dialect baseline described in §4.1 (here referred to as the **NorSID expert**), as well as three Norwegian language experts.

To produce the language experts, we fine-tune[11] with the masked language modeling (MLM) objective using an example from Sentence Transformers

[7]The dialect distribution of this treebank is different than that of NoMusic, with around 30% each of East, North, and West Norwegian sentences, and 7% Trønder. We use the script by Blaschke et al. (2023) to merge the phonetic transcriptions with the treebank.

[8]We merge the two geo-political entity types `GPE_LOC` and `GPE_ORG` into location and organization, respectively, and entirely remove the category of miscellaneous entities, since it occurs very rarely in the dataset.

[9]`mainlp_slots3_scandibert_deprel_sid_8446`

[10]`https://github.com/monologg/JointBERT` (Apache 2.0; commit `00324f6`)

[11]mDeBERTa is trained using replaced token detection (RTD; He et al., 2023) rather than MLM, hence we do not consider MLM fine-tuning a continuation of training.

(Reimers and Gurevych, 2019).[12] We train for 20 epochs and select the best checkpoint based on perplexity on the development set of NoMusic.

We use three different datasets for the language experts to examine whether the text style/genre and language variety makes a difference: the Bokmål transcriptions of interviews in the Nordic Dialect Corpus (NDC; Johannessen et al., 2009),[13] the Bokmål part of the Norwegian Dependency Treebank (NDT; Solberg et al., 2014) which contains news articles, blog posts, and government reports/transcripts,[14] and the NoMusic training set.

Identifying layers to replace As an ablation experiment to identify layers of the EnSID expert that might be replaceable, we revert its layers back to their state in the pretrained model and observe the performance of the resulting model on the NoMusic development set. For each of the mDeBERTa models fine-tuned on the English xSID data with three different seeds, we revert pairs of sequential layers (i.e., 0,1, then 1,2, and so on).

Unlike Bandarkar et al. (2024), we are unable to use the mean absolute value (MAV) of the difference in parameters through fine-tuning to identify less salient layers. The variance in change of the parameters of the EnSID expert is very small at 1.5×10^{-7}, such that no layers exhibit significantly higher MAVs than others. This may be due to any number of differences of our setup, such as model architecture, layer depth, fine-tuning objective, amount of fine-tuning data, or simply duration of fine-tuning; further analysis of layer-wise training dynamics is left to future work.

Model assembly The layer-reverting experiments identify the first two layers of the EnSID expert as suited for replacement. We replace the token embeddings and the first two encoder layers of the EnSID expert with the corresponding layers of the Norwegian expert, resulting in four assembled models, one for each Norwegian expert. We do not merge any parameters.

Shared task submission We submit the model produced by assembling layers of the NorSID expert and the EnSID expert.[15]

[12] https://github.com/UKPLab/sentence-transformers (Apache 2.0; commit 1cb196a)

[13] We use a random 80:10:10 split of half of the corpus.

[14] This treebank is the basis of the NorNE dataset (§4.3).

[15] mainlp_{slots,intents}2_mdeberta_topline_swapped

Training data	Model	Intents	Slots
English (train)	NorBERT	95.1 $_{0.2}$	79.7 $_{0.4}$
	ScandiBERT	94.8 $_{0.8}$	80.7 $_{0.7}$
	mDeBERTa	92.4 $_{1.8}$	76.5 $_{1.2}$
Norwegian (train, MT)	NorBERT	96.2 $_{0.5}$	53.9 $_{0.3}$
	ScandiBERT	96.3 $_{0.1}$	54.6 $_{0.4}$
	mDeBERTa	96.7 $_{0.3}$	55.2 $_{1.1}$
Nor. dialect (dev, 90%)	NorBERT	94.2 $_{0.6}$	76.8 $_{1.1}$
	ScandiBERT	92.8 $_{0.6}$	81.2 $_{0.6}$
	mDeBERTa	93.4 $_{0.7}$	83.2 $_{1.0}$ *

Table 2: **Test scores of baseline models** (intent accuracy in %, slot span F_1 in %) trained on English data, machine-translated Norwegian data, or 90% of the dialectal Norwegian development set. The results are averaged over three runs, with standard deviations as subscripts. * Model submitted to the shared task (slots and intents).

5 Results and Analysis

In this section, we mainly focus on the test scores. For the shared task, we submitted models considering their development set performance. These are denoted by asterisks in the results tables, and further discussed in §5.5. All models were trained (and evaluated on the development set) before the test set was released.

Table 8 in Appendix B shows the development and test scores for all systems.

5.1 Baselines

The training data choice had a greater effect on the SID quality than the PLM choice (Table 2).

Training data Despite the language difference, the models trained on the English training data provide strong baselines – especially for the Norwegian and Scandinavian PLMs, which achieve intent prediction accuracies of 94.8–95.1% and slot-filling F_1 scores of 79.7–80.7%.

The models trained on the machine-translated Norwegian training set produce better intent labels (with accuracies between 96.2 and 94.8%), but are poor slot fillers (53.9–55.2% F_1).[16] We suspect this is due to quality issues with the slot label

[16] This is similar to the results by (van der Goot et al., 2021a), who find training on translated data to be beneficial for intent classification. In their experiments, translated data improves slot filling for a PLM with poor baseline scores for cross-lingual slot filling, but lowers the performance of another model whose cross-lingual slot-filling scores were already quite high when trained on English data.

projections. To substantiate this, we compare the strict span F_1 scores with their loose counterpart, which allows spans that only partially overlap. Although the models trained on machine-translated Norwegian achieve much lower strict F_1 scores, the loose F_1 scores are similar to those of the other baselines (Table 9, Appendix B). This suggests that the slot annotations of the machine-translated Norwegian baselines mainly suffer from incorrect spans, as would be expected from poor projections, which affect the span, but not the label.

Training the models on the largely dialectal development set led to overfitting – these models show the greatest drop between development and test set performance (Table 8 in Appendix B). This may have been exacerbated by how we stratified the data, as we did not ensure that all translations of the same sentence were assigned to the same split. Furthermore, the development set is significantly smaller than the training set (2.9k vs. 43.6k samples). Finally, one intent and one slot type were present in the test but not in the development set, as well as seven `I` labels (though the corresponding `B` was seen, more on this under Limitations). Despite all of this, the models fine-tuned on this dataset produce some of the best slot annotations (with F_1 scores between 76.8 and 83.2%).

PLM No PLM is consistently the best or worst model. For the models trained on the English or machine-translated Norwegian data, performance on slot filling appears to be correlated with performance on intent classification, and vice versa. However, there seems to be no relation between the two for the models trained on the dialectal data where, e.g., NorBERT produces the best intent labels but the worst slot annotations.

5.2 Character-Level Noise

Fine-tuning on noised data generally improves the models' performance (Table 3) – by up to 1.2 percentage points (pp.) for intent classification and up to 1.3 pp. for slot filling. Which noise level helps most depends on the PLM choice; this is similar to previous findings on using noised data for POS tagging in Norwegian dialects and other language varieties (Blaschke et al., 2023). However, the effect of noise also depends on the task – the trends are different for intent classification and slot filling.

Prior work has found the ratios of words that were split into multiple subword tokens to be a strong predictor for transfer success between closely related varieties: the more similar the split word ratios are in the training and evaluation data, the more successful transfer tends to be (Blaschke et al., 2023). In our study, only the intent classification results correlate with this difference in split word ratio (Table 10 in Appendix B). We hypothesize that the weak correlations with the slot-filling results might be due to the mixed quality of the silver-standard slot annotations in the training data.

PLM	Noise (%)	Intents	Δ	Slots	Δ	
NorBERT	0	96.2 $_{0.5}$		53.9 $_{0.3}$		
	10	96.4 $_{0.3}$	+0.2	55.1 $_{0.8}$	+1.2	
	20	97.2 $_{0.2}$	+1.0	55.0 $_{0.3}$	+1.1	
	30	97.4 $_{0.5}$	+1.2	54.1 $_{1.1}$	+0.1	
ScandiBERT	0	96.3 $_{0.1}$		54.6 $_{0.4}$		
	10	96.5 $_{0.4}$	+0.2	55.9 $_{0.7}$	+1.3	
	20	97.5 $_{0.2}$	+1.2	54.6 $_{0.5}$	–0.0	
	30	97.1 $_{0.5}$	+0.8	54.8 $_{0.5}$	+0.2	
mDeBERTa	0	96.7 $_{0.3}$		55.2 $_{1.1}$		
	10	96.5 $_{0.9}$	–0.2	55.6 $_{1.0}$	+0.4	
	20	97.5 $_{0.2}$	+0.8	55.5 $_{0.5}$	+0.2	*
	30	97.0 $_{0.5}$	+0.3	56.2 $_{0.5}$	+1.0	

Table 3: **Test scores of models trained on noised data** (intent accuracy in %, slot span F_1 in %). The results are averaged over three runs, with standard deviations as subscripts. The Δ columns show the differences to the respective baseline (0 % noise). * Model submitted to the shared task (intents).

5.3 Auxiliary Tasks

The effect of the auxiliary tasks depends on the tasks themselves, the language of the SID data, and whether they are trained before or simultaneously with the target SID task. Table 4 shows the results on the SID test data; Table 11 in Appendix B also shows the development scores on the SID and auxiliary task data.

Intermediate-task training vs. multi-task learning For slot filling, intermediate-task training (training first on the auxiliary task and afterwards on the SID data) generally achieves better results than simultaneous multi-task learning. For intent classification, there is no clear trend.

We additionally examine whether the effects of multi-task learning are similar across tasks by inspecting the models' performances on the development sets of the auxiliary tasks (Table 11 in Appendix B). For the auxiliary tasks, multi-task learning nearly always yields worse results than

Task		Intents	Δ	Slots	Δ	
English training data						
Baseline		94.8 $_{0.8}$		80.7 $_{0.7}$		
Dial	×	83.8 $_{3.2}$	–11.0	75.8 $_{1.2}$	–4.9	
	→	94.0 $_{1.7}$	–0.8	79.2 $_{0.9}$	–1.5	
POS	×	94.9 $_{0.3}$	+0.0	81.1 $_{0.3}$	+0.4	
	→	94.7 $_{0.2}$	–0.2	82.2 $_{1.1}$	+1.5	
Dep	×	93.5 $_{0.6}$	–1.3	81.5 $_{0.2}$	+0.8	
	→	94.9 $_{1.2}$	+0.1	81.8 $_{0.7}$	+1.1	*
NER	×	95.3 $_{1.0}$	+0.5	80.6 $_{1.0}$	–0.1	
	→	95.0 $_{0.4}$	+0.1	81.1 $_{0.9}$	+0.4	
Machine-translated Norwegian training data						
Baseline		96.3 $_{0.1}$		54.6 $_{0.4}$		
Dial	×	89.2 $_{1.4}$	–7.1	51.7 $_{0.1}$	–2.9	
	→	95.2 $_{1.0}$	–1.1	53.7 $_{0.5}$	–0.9	
POS	×	96.8 $_{0.4}$	+0.4	53.7 $_{0.6}$	–0.9	
	→	96.7 $_{0.4}$	+0.3	54.4 $_{0.8}$	–0.2	
Dep	×	96.9 $_{0.3}$	+0.5	53.7 $_{0.2}$	–0.9	
	→	96.4 $_{0.3}$	+0.1	54.8 $_{0.6}$	+0.2	
NER	×	96.9 $_{0.1}$	+0.6	53.8 $_{0.3}$	–0.8	
	→	96.4 $_{0.5}$	+0.1	53.5 $_{1.0}$	–1.1	

Table 4: **Test scores of models trained on auxiliary tasks** (intent accuracy in %, slot span F_1 in %). The results are averaged over three runs, with standard deviations as subscripts. The Δ columns show the differences to the respective baseline. Key: *Dial* = dialect identification, *dep* = dependency parsing, × = multitask learning, → = intermediate-task training. * Model submitted to the shared task (slots).

exclusively training on the auxiliary tasks (as the first step in intermediary-task training). Although the performance gap between the two settings is especially large for the two syntactic tasks (with multi-task learning achieving scores that are 11.8–26.7 pp. lower), the impact on the corresponding SID performance is less clear-cut (with multi-task learning leading by up to 0.5 pp. in some constellations and falling behind by 2.1 pp. in others).

Auxiliary task choice and SID training language Dialect identification diminishes both the intent classification and slot-filling performance in all of our set-ups (most drastically in the multi-task set-up with the English SID data, with drops of 11.0 pp. for intent classification and 4.9 pp. for slot filling).

The effects of the other tasks depend on the SID training language. For the models fine-tuned on Norwegian data, the other tasks slightly improve intent classification performance (with gains of up to 0.6 pp.) but typically negatively impact slot filling (with changes between +0.2 and –0.9 pp.) – the grammatical tasks do not mitigate the effect of poor slot annotations in the machine-translated data.

For the English SID training data, the syntax-related tasks (POS tagging and dependency parsing) improve slot filling by between 0.4 and 1.5 pp., but have no or a negative effect on the intent classification performance (changes to the baseline between +0.1 and –1.3 pp.). Despite positive prior findings (Krückl et al., 2025), NER has no or only slightly positive effects on either SID task.

Dialects There is no apparent connection between the dialect distributions in the auxiliary task training data and the SID performance on the different dialect groups (Table 12 in Appendix B). This applies both to the models trained on English SID data and on the Norwegian translations, although the gains per dialect group differ between them.

For the syntactic tasks, one possible explanation is that the dialect transcriptions do not sufficiently align with the ad-hoc dialect spellings used in NoMusic to show strong effects based on the represented dialect groups.

5.4 Layer Swapping

Identifying layers to replace Results of reverting pairs of layers of the EnSID expert are shown in Table 5. We found that in general, performance decreased as later layers were reverted. This aligns with our intuition that the later layers, being closer to the classification heads, are particularly important for performance.

Notably, we found that reverting layers 0 and 1 slightly increased performance on both slot filling and intent classification (across three runs we observed an average improvement of slot F_1 of 3.0 pp. and intent accuracy of 0.9 pp.). This improvement through reverting is somewhat surprising, and suggests that something about the fine-tuning process on the English data is counterproductive to the robustness of the model to out-of-language data, at least where the first two layers are concerned.

We also observed a large variance in the effect of reverting the last two layers on intent classification, this is due to the first seed seeing quite a large drop (to 55.7%, the average accuracy of the other two seeds was 86.1%).

Layers	Intents	Δ	Slots	Δ
none	95.1 $_{0.6}$		77.1 $_{1.4}$	
0,1	96.1 $_{0.4}$	+0.9	80.1 $_{0.6}$	+3.0
1,2	96.0 $_{0.2}$	+0.9	77.8 $_{1.0}$	+0.6
2,3	95.1 $_{0.8}$	0.0	69.9 $_{0.5}$	–7.2
3,4	93.9 $_{0.6}$	–1.2	65.8 $_{0.2}$	–11.3
4,5	93.8 $_{0.7}$	–1.4	68.0 $_{1.6}$	–9.1
5,6	93.6 $_{1.1}$	–1.5	69.3 $_{2.0}$	–7.8
6,7	90.7 $_{0.9}$	–4.5	63.1 $_{5.8}$	–14.0
7,8	87.0 $_{1.2}$	–8.1	59.7 $_{3.8}$	–17.4
8,9	84.3 $_{2.7}$	–10.8	58.0 $_{2.5}$	–19.1
9,10	72.6 $_{9.4}$	–22.6	54.1 $_{5.5}$	–23.0
10,11	76.0 $_{17.9}$	–19.2	59.4 $_{0.6}$	–17.7

Table 5: **Development scores of the EnSID expert with reverted layers** (intent accuracy in %, slot span F_1 in %). The results are averaged over three runs with standard deviations as subscripts.

Norwegian Expert	Intents	Δ	Slots	Δ
N/A – EnSID unchanged	95.1 $_{0.6}$		77.1 $_{1.4}$	
N/A – EnSID reverted (0,1)	96.1 $_{0.4}$	+0.9	80.1 $_{0.6}$	+3.0
NorSID expert	98.3 $_{0.4}$	+2.2	86.5 $_{0.6}$	+9.6
NoMusic MLM	96.9	+0.8	78.8	+1.7
NDT MLM	97.4	+1.3	78.6	+1.5
NDC MLM	96.3	+0.2	77.9	+0.8

Table 6: **Development scores of *assembled* models using different Norwegian experts** (intent accuracy in %, slot span F_1 in %). Each Norwegian expert is assembled with the EnSID expert. Results for the unchanged EnSID expert and the best reverted model, layers 0,1, are shown for comparison, each averaged over three runs. The assembled model with the NorSID expert is averaged over nine runs (for each combination of NorSID and EnSID expert). We don't repeat runs for unpromising language experts. The standard deviation, where applicable, is denoted by subscripts.

Choosing a complementary expert Table 6 shows the results of replacing the first two layers of the EnSID expert with the corresponding layers of each of our Norwegian experts. These combinations performed roughly on par with or slightly better than the reverted model, except for the model containing the layers from the NorSID expert, which performed better, particularly for slot filling. Further analysis is needed to better understand what makes layers useful for assembling into a model, this is left for future work.

As these were exploratory preliminary experiments, we do not repeat runs for unpromising language experts.

	Intents		Slots	
	dev (no)	test (no)	dev (no)	test (no)
EnSID expert	95.1 $_{0.6}$	92.0 $_{0.8}$	78.6 $_{1.1}$	77.2 $_{1.6}$
NorSID expert	**99.4** $_{0.0}$	93.4 $_{0.7}$	**96.4** $_{0.4}$	83.2 $_{1.0}$
Assembled*	98.3 $_{0.4}$	**96.4** $_{0.2}$	86.5 $_{0.6}$	**84.9** $_{0.5}$
	dev (en)	test (en)	dev (en)	test (en)
EnSID expert	**100.0** $_{0.0}$	99.2 $_{0.0}$	97.1 $_{0.3}$	**96.0** $_{0.3}$
NorSID expert	**100.0** $_{0.0}$	**100.0** $_{0.0}$	90.1 $_{1.0}$	80.9 $_{1.4}$
Assembled*	**100.0** $_{0.0}$	99.3 $_{0.2}$	**97.5** $_{0.3}$	**96.0** $_{0.3}$

Table 7: **Development and test scores of the original experts and assembled model on NoMusic (no) and xSID 0.6 English (en)** (intent accuracy and slot F_1 in %, best results bolded). The results are averaged over three runs for the experts, and over nine runs for the assembled model, with standard deviations as subscripts. * Model submitted to the shared task (slots and intents).

Final submission Results for the submitted assembled model (layers from the EnSID and NorSID expert), and the individual experts on both NoMusic and the xSID 0.6 English set are shown in Table 7. Overall, the assembled model is more robust to out-of-language data than the respective experts, outperforming the EnSID expert on the Norwegian development and test sets, and mostly outperforming the Norwegian expert on the English sets, except for intent classification on the test set. We hypothesize that this exception may be due to the EnSID expert overfitting the intent classification task, which was not mitigated by using the first two layers of the Norwegian SID expert.

Using only two layers of the Norwegian SID expert, which suffered from overfitting (§5.1), seems to have a regularizing effect, as the assembled model outperforms the Norwegian SID expert in both tasks on the Norwegian test set.

5.5 Results of Shared Task Submissions

We submitted three systems per task (slot and intent detection) and did not participate in dialect classification. The official results are provided in the shared task overview paper (Scherrer et al., 2025) and the accompanying website,[17] and we include them in Table 8 (Appendix B). Unlike the previous sections, they only represent a single random seed. Of our intent classification systems, noise injection worked best (ranked 5th of all submissions; 97.64% accuracy), narrowly followed by Layer Swapping

[17] https://github.com/ltgoslo/NoMusic/blob/main/NorSID/results.md

(6th rank; 97.16%). Both beat the baseline trained only on the dialectal development set (10th rank; 93.47%).

For slot detection, Layer Swapping instead was our best method, ranking third in the competition (85.57% F_1). Compared to our other two submissions – the baseline trained on the development set (5th rank; 83.68%) and the model with intermediate-task training on dependency parsing (6th rank; 82.57%) – it performed best on three out of the four Norwegian varieties.

6 Discussion and Conclusion

The strength of our baselines suggest that the NorSID task is, relatively speaking, less challenging than other dialectical variants of xSID (cf. van der Goot et al., 2021a; Aepli et al., 2023; Srivastava and Chiang, 2023b; Kwon et al., 2023; Winkler et al., 2024; Muñoz-Ortiz et al., 2025; Krückl et al., 2025). We suspect that there is less deviation from standard Norwegian, and less variation between the dialects. This limits the gains we could expect from additional methods, particularly on the intent classification task, where the accuracy of our baselines ranges from 92.4% to 96.7% on the test set.

We observe somewhat of a trade-off between performance on intent classification (strongest for models trained on Norwegian data) and slot filling (strongest for models trained on the gold-standard English training or Norwegian development data; §5.1). We hypothesize that the latter is due to the poor quality of the slot labels in the machine-translated Norwegian training data.

We see noise injection as a simple way to improve transfer between a standard language and related varieties (§5.2), although it requires access to appropriate training data. Where a language has enough resources for additional annotated datasets, we see mixed effects from the inclusion of auxiliary NLP tasks (§5.3). Which auxiliary tasks help SID performance depends on the target-task training data and SID subtask (intent classification vs. slot filling) and remains hard to predict, requiring further research.

Improving performance on the slot-filling task proved to be quite difficult; our most successful method by a small margin is the assembled model made up of layers from a model trained on the NoMusic development set (NorSID expert), and another on the English xSID data (§5.4). Using layers from both of these models seems to have a regularizing effect and produces a model that is able to perform well on both languages and suffers less from overfitting than the NorSID expert.

We successfully adapted Layer Swapping – originally applied to a 32-layer decoder – to a 12-layer encoder, demonstrating its potential for resource-efficient cross-lingual transfer. Layer Swapping could prove useful for modular solutions, as layers for different languages could dynamically replace those of a "base" SID expert to adapt the model. We again note that the subset of the development set of NoMusic we used, at 2.9k examples, is much smaller than the set used to train our EnSID expert, at 43.6k examples; this modular approach would allow adaptation to different languages in a fairly lightweight manner post-hoc.

We encourage further research comparing (and combining) different methods for low-resource NLP with the same training and/or evaluation data.

Limitations

Both MaChAmp and the JointBERT implementation only consider the exact labels seen during training; consequently our SID models will not predict unseen `I` tags, even if the corresponding `B` tag is known. In particular, the English xSID sets have fewer `I` tags, i.e., corresponding slots are sometimes spread over more words in NoMusic. We also find that the NoMusic test set has more `I` tags than the development set.

While we compare several different approaches for improving SID on this task, we find the conditions of their success are difficult to generalize. For example, no auxiliary task has prevailed. For Layer Swapping, it is not clear what makes layers of particular expert suitable for assembly, and whether our findings generalize to other models, languages, or tasks. Further work is needed to understand which method will work best for what conditions, and how best to apply each method.

Because of time constraints, we were not able to further investigate the effect of including auxiliary datasets in standard vs. dialectal varieties. In particular, it would be interesting to include POS tagging and dependency parsing on Bokmål or Nynorsk data (e.g., the NDT and LIA treebanks we used in other ways in this paper).

Similarly, we did not try MLM fine-tuning using the dialect version of NDC to produce an expert for Layer Swapping; on inspection of the corpus, the Bokmål version seemed closer to the target language, and given the unpromising results using the other MLM experts we did not explore this further.

Acknowledgements

We thank the anonymous reviewers for their suggestions. This research is supported by European Research Council (ERC) Consolidator Grant DIALECT 101043235.

References

Khadige Abboud and Gokmen Oz. 2024. Towards equitable natural language understanding systems for dialectal cohorts: Debiasing training data. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16487–16499, Torino, Italia. ELRA and ICCL.

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial evaluation campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.

Noëmi Aepli and Rico Sennrich. 2022. Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4074–4083, Dublin, Ireland. Association for Computational Linguistics.

Ekaterina Artemova, Verena Blaschke, and Barbara Plank. 2024. Exploring the robustness of task-oriented dialogue systems for colloquial German varieties. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 445–468, St. Julian's, Malta. Association for Computational Linguistics.

Lucas Bandarkar, Benjamin Muller, Pritish Yuvraj, Rui Hou, Nayan Singhal, Hongjiang Lv, and Bing Liu. 2024. Layer swapping for zero-shot cross-lingual transfer in large language models. *Preprint*, arXiv:2410.01335.

Verena Blaschke, Barbara Kovačić, Siyao Peng, Hinrich Schütze, and Barbara Plank. 2024. MaiBaam: A multi-dialectal Bavarian Universal Dependency treebank. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10921–10938, Torino, Italia. ELRA and ICCL.

Verena Blaschke, Hinrich Schütze, and Barbara Plank. 2023. Does manipulating tokenization aid cross-lingual transfer? a study on POS tagging for non-standardized languages. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 40–54, Dubrovnik, Croatia. Association for Computational Linguistics.

Maharaj Brahma, Kaushal Maurya, and Maunendra Desarkar. 2023. SelectNoise: Unsupervised noise injection to enable zero-shot machine translation for extremely low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1615–1629, Singapore. Association for Computational Linguistics.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. BERT for joint intent classification and slot filling. *Preprint*, arXiv:1902.10909.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: An embedded spoken language understanding system for private-by-design voice interfaces. *Preprint*, arXiv:1805.10190.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Hossam Elkordi, Ahmed Sakr, Marwan Torki, and Nagwa El-Makky. 2024. AlexuNLP24 at AraFinNLP2024: Multi-dialect Arabic intent detection with contrastive learning in banking domain. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 415–421, Bangkok, Thailand. Association for Computational Linguistics.

Murhaf Fares and Samia Touileb. 2024. BabelBot at AraFinNLP2024: Fine-tuning t5 for multi-dialect intent detection with synthetic data and model ensembling. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 433–440, Bangkok, Thailand. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kristin Hagen, Live Håberg, Eirik Olsen, and Åshild Søfteland. 2018. Transkripsjonsrettleiing for LIA.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Janne Bondi Johannessen, Joel James Priestley, Kristin Hagen, Tor Anders Åfarli, and Øystein Alexander Vangsnes. 2009. The Nordic dialect corpus–an advanced research tool. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, pages 73–80, Odense, Denmark. Northern European Association for Language Technology (NEALT).

Fredrik Jørgensen, Tobias Aasmoe, Anne-Stine Ruud Husevåg, Lilja Øvrelid, and Erik Velldal. 2020. NorNE: Annotating named entities for Norwegian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4547–4556, Marseille, France. European Language Resources Association.

Xaver Maria Krückl, Verena Blaschke, and Barbara Plank. 2025. Improving dialectal slot and intent detection with auxiliary tasks: A multi-dialectal Bavarian case study. In *Proceedings of the Twelfth Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Sang Yun Kwon, Gagan Bhatia, Elmoatez Billah Nagoudi, Alcides Alcoba Inciarte, and Muhammad Abdul-mageed. 2023. SIDLR: Slot and intent detection models for low-resource language varieties. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 241–250, Dubrovnik, Croatia. Association for Computational Linguistics.

John Lafferty, Andrew McCallum, Fernando Pereira, et al. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962, Online. Association for Computational Linguistics.

Petter Mæhlum and Yves Scherrer. 2024. NoMusic - the Norwegian multi-dialectal slot and intent detection corpus. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 107–116, Mexico City, Mexico. Association for Computational Linguistics.

Sanad Malaysha, Mo El-Haj, Saad Ezzini, Mohammed Khalilia, Mustafa Jarrar, Sultan Almujaiwel, Ismail Berrada, and Houda Bouamor. 2024. AraFinNLP 2024: The first Arabic financial NLP shared task. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 393–402, Bangkok, Thailand. Association for Computational Linguistics.

Syrielle Montariol, Arij Riabi, and Djamé Seddah. 2022. Multilingual auxiliary tasks training: Bridging the gap between languages for zero-shot transfer of hate speech detection models. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 347–363, Online only. Association for Computational Linguistics.

Alberto Muñoz-Ortiz, Verena Blaschke, and Barbara Plank. 2025. Evaluating pixel language models on non-standardized languages. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Lilja Øvrelid, Andre Kåsen, Kristin Hagen, Anders Nøklestad, Per Erik Solberg, and Janne Bondi Johannessen. 2018. The LIA treebank of spoken Norwegian dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Vishakh Padmakumar, Leonard Lausen, Miguel Ballesteros, Sheng Zha, He He, and George Karypis. 2022. Exploring the role of task transferability in large-scale multi-task learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2542–2550, Seattle, United States. Association for Computational Linguistics.

Clifton Poth, Jonas Pfeiffer, Andreas Rücklé, and Iryna Gurevych. 2021. What to pre-train on? Efficient intermediate task selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10585–10605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

Asmaa Ramadan, Manar Amr, Marwan Torki, and Nagwa El-Makky. 2024. MA at AraFinNLP2024: BERT-based ensemble for cross-dialectal Arabic intent detection. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 441–445, Bangkok, Thailand. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *Preprint*, arXiv:1706.05098.

David Samuel, Andrey Kutuzov, Samia Touileb, Erik Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel, and Anna Palatkina. 2023. NorBench – a benchmark for Norwegian language models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 618–633, Tórshavn, Faroe Islands. University of Tartu Library.

Yves Scherrer, Rob van der Goot, and Petter Mæhlum. 2025. VarDial evaluation campaign 2025: Norwegian slot and intent detection and dialect identification (NorSID). In *Proceedings of the Twelfth Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Fynn Schröder and Chris Biemann. 2020. Estimating the influence of auxiliary tasks for multi-task learning of sequence tagging tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2971–2985, Online. Association for Computational Linguistics.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. Transfer to a low-resource language via close relatives: The case study on Faroese. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 728–737, Tórshavn, Faroe Islands. University of Tartu Library.

Per Erik Solberg, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen, and Janne Bondi Johannessen. 2014. The Norwegian dependency treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 789–795, Reykjavik, Iceland. European Language Resources Association (ELRA).

Aarohi Srivastava and David Chiang. 2023a. BERTwich: Extending BERT's capabilities to model dialectal and noisy text. In *Findings of the Association for Computational Linguistics: EMNLP*

2023, pages 15510–15521, Singapore. Association for Computational Linguistics.

Aarohi Srivastava and David Chiang. 2023b. Fine-tuning BERT with character-level noise for zero-shot transfer to dialects and closely-related languages. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 152–162, Dubrovnik, Croatia. Association for Computational Linguistics.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021a. From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2479–2497, Online. Association for Computational Linguistics.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021b. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Miriam Winkler, Virginija Juozapaityte, Rob van der Goot, and Barbara Plank. 2024. Slot and intent detection resources for Bavarian and Lithuanian: Assessing translations vs natural queries to digital assistants. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14898–14915, Torino, Italia. ELRA and ICCL.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online. Association for Computational Linguistics.

Kaisheng Yao, Geoffrey Zweig, Mei-Yuh Hwang, Yangyang Shi, and Dong Yu. 2013. Recurrent neural networks for language understanding. In *Interspeech 2013*, pages 2524–2528.

A Spelling Changes to the Dialectological Transcriptions

We make slight changes to the dialectological transcriptions used in LIA based on LIA's transcription guidelines (Hagen et al., 2018). The idea is to turn the transcriptions into slightly more plausible spellings, but we want to stress that these rules are simplistic and not meant to produce text that fully emulates naturalistic dialect spellings.

- We replace ⟨`L`⟩ (/ɽ/, *tjukk l* 'thick l') with ⟨`l`⟩. While it can also correspond to ⟨`rd`⟩, we found that it much more often corresponds to ⟨`l`⟩ in the data.
- We remove apostrophes (originally used to mark syllabic consonants).
- The dialectological transcriptions use double consonants to mark short vowels, which can lead to consonant clusters that are unlikely to occur in written Norwegian. In words where a double consonant is followed by at least one more consonant, we remove one of the doubled consonants ($C_1C_1C_2 \rightarrow C_1C_2$). If the sequence is ⟨`ssjt`⟩ or ⟨`ssjk`⟩, we instead replace it with ⟨`rst`⟩ or ⟨`rsk`⟩, respectively. If it otherwise starts with ⟨`ssj`⟩ or ⟨`kkj`⟩, we do not remove the first ⟨`s`⟩ or ⟨`k`⟩.

B Detailed Results

All Table 8 shows the development and test scores of all models (described throughout §5).

Baselines Table 9 provides the results of additional slot-filling metrics for the baselines (§5.1).

Noise Table 10 shows the correlations between the split word ratio difference of the noised training sets and the dialectal evaluation sets (cf. §5.2).

Auxiliary tasks The remaining tables provide additional details for §5.3. Table 11 focuses on the set-ups with auxiliary tasks and shows the scores on these tasks in addition to the SID scores. Table 12 focuses on the models trained on auxiliary tasks and shows the dialect distributions in the auxiliary task training data as well as the dialect-wise SID results.

Training data	PLM	Details	Intents (acc., %)			Slots (span F_1, %)		
			Dev	Test	Subm.	Dev	Test	Subm.
English (train)	NorBERT	baseline	$96.9_{0.4}$	$95.1_{0.2}$		$79.9_{0.1}$	$79.7_{0.4}$	
	ScandiBERT	baseline	$96.4_{0.5}$	$94.8_{0.8}$		$81.3_{0.3}$	$80.7_{0.7}$	
		dial ×	$84.3_{2.7}$	$83.8_{3.2}$		$75.8_{1.8}$	$75.8_{1.2}$	
		dial →	$95.8_{0.7}$	$94.0_{1.7}$		$79.7_{1.3}$	$79.2_{0.9}$	
		POS ×	$96.0_{0.5}$	$94.9_{0.3}$		$81.6_{0.2}$	$81.1_{0.3}$	
		POS →	$96.3_{0.2}$	$94.7_{0.2}$		$82.3_{0.8}$	$82.2_{1.1}$	
		dep ×	$94.7_{1.2}$	$93.5_{0.6}$		$82.0_{0.4}$	$81.5_{0.2}$	
		dep →	$96.7_{1.0}$	$94.9_{1.2}$		$82.5_{0.9}$	$81.8_{0.7}$	82.6
		NER ×	$97.2_{0.9}$	$95.3_{1.0}$		$80.9_{0.9}$	$80.6_{1.0}$	
		NER →	$96.8_{0.4}$	$95.0_{0.4}$		$81.3_{1.1}$	$81.1_{0.9}$	
	mDeBERTa	baseline	$95.3_{1.1}$	$92.4_{1.8}$		$77.3_{1.2}$	$76.5_{1.2}$	
Norwegian (MT) (train)	NorBERT	baseline	$98.3_{0.4}$	$96.2_{0.5}$		$55.7_{0.5}$	$53.9_{0.3}$	
		noise (10%)	$98.5_{0.4}$	$96.4_{0.3}$		$57.3_{0.2}$	$55.1_{0.8}$	
		noise (20%)	$98.6_{0.2}$	$97.2_{0.2}$		$56.4_{0.4}$	$55.0_{0.3}$	
		noise (30%)	$99.0_{0.1}$	$97.4_{0.5}$		$56.4_{1.3}$	$54.1_{1.1}$	
	ScandiBERT	baseline	$97.6_{0.0}$	$96.3_{0.1}$		$55.5_{0.4}$	$54.6_{0.4}$	
		noise (10%)	$97.8_{0.1}$	$96.5_{0.4}$		$56.9_{0.9}$	$55.9_{0.7}$	
		noise (20%)	$98.4_{0.3}$	$97.5_{0.2}$		$55.6_{0.8}$	$54.6_{0.5}$	
		noise (30%)	$98.0_{0.4}$	$97.1_{0.5}$		$56.5_{0.6}$	$54.8_{0.5}$	
		dial ×	$89.8_{1.4}$	$89.2_{1.4}$		$53.0_{0.1}$	$51.7_{0.1}$	
		dial →	$96.1_{1.0}$	$95.2_{1.0}$		$54.4_{0.4}$	$53.7_{0.5}$	
		POS ×	$97.9_{0.3}$	$96.8_{0.4}$		$54.0_{0.5}$	$53.7_{0.6}$	
		POS →	$97.8_{0.5}$	$96.7_{0.4}$		$55.5_{0.4}$	$54.4_{0.8}$	
		dep ×	$98.0_{0.4}$	$96.9_{0.3}$		$54.5_{0.2}$	$53.7_{0.2}$	
		dep →	$97.5_{0.7}$	$96.4_{0.3}$		$55.7_{0.5}$	$54.8_{0.6}$	
		NER ×	$97.9_{0.6}$	$96.9_{0.1}$		$54.6_{0.5}$	$53.8_{0.3}$	
		NER →	$97.6_{0.2}$	$96.4_{0.5}$		$54.1_{0.6}$	$53.5_{1.0}$	
	mDeBERTa	baseline	$98.4_{0.4}$	$96.7_{0.3}$		$56.5_{0.2}$	$55.2_{1.1}$	
		noise (10%)	$98.5_{0.6}$	$96.5_{0.9}$		$56.7_{0.7}$	$55.6_{1.0}$	
		noise (20%)	$99.2_{0.1}$	$97.5_{0.2}$	97.6	$56.4_{0.2}$	$55.5_{0.5}$	
		noise (30%)	$98.9_{0.5}$	$97.0_{0.5}$		$57.6_{0.3}$	$56.2_{0.5}$	
Nor. dialect (dev 90%)	NorBERT	baseline[1]	$99.4_{0.0}$	$94.2_{0.6}$		$94.5_{0.6}$	$76.8_{1.1}$	
	ScandiBERT	baseline[1]	$99.6_{0.2}$	$92.8_{0.6}$		$96.0_{0.6}$	$81.2_{0.6}$	
	mDeBERTa	baseline[1]	$99.4_{0.0}$	$93.4_{0.7}$	93.5	$96.4_{0.4}$	$83.2_{1.0}$	83.7
Nor. dialect (dev 90%) / English	mDeBERTa	EnSID expert	$95.1_{0.6}$	$92.0_{0.8}$		$78.6_{1.1}$	$77.2_{1.6}$	
		NorSID expert[1,2]	$99.4_{0.0}$	$93.4_{0.7}$		$96.4_{0.4}$	$83.2_{1.0}$	
		assembled	$98.3_{0.4}$	$96.4_{0.2}$	97.2	$86.5_{0.6}$	$84.9_{0.5}$	85.6

Table 8: **Intent classification and slot-filling scores for all systems** on the development and test data, and for the runs we submitted to the shared task. Results are averaged across three runs, with the exception of the assembled system, which is averaged across nine total combinations of three runs each of both experts. Standard deviations are denoted by subscripts. Key: *Dial* = dialect identification, *dep* = dependency parsing, × = multitask learning, → = intermediate-task training. [1]For the models trained on 90% of the development data, the dev scores are measured on the remaining 10%. [2]The results for this model are already listed in the Norwegian dialect section (mDeBERTa), but repeated here for easier comparison.

Training data	Model	Loose	Unlabelled	Strict
English (train)	NorBERT	84.4 $_{1.0}$	84.4 $_{1.3}$	76.5 $_{1.2}$
	ScandiBERT	88.0 $_{0.3}$	88.2 $_{0.2}$	80.7 $_{0.7}$
	mDeBERTa	86.8 $_{0.2}$	87.0 $_{0.7}$	79.7 $_{0.4}$
Norwegian (MT) (train)	NorBERT	84.4 $_{0.5}$	63.4 $_{0.7}$	53.9 $_{0.3}$
	ScandiBERT	86.0 $_{0.6}$	62.9 $_{0.4}$	54.6 $_{0.4}$
	mDeBERTa	85.7 $_{0.4}$	63.3 $_{1.1}$	55.2 $_{1.1}$
Nor. dialect (dev 90%)	NorBERT	84.9 $_{1.0}$	90.5 $_{0.2}$	76.8 $_{1.1}$
	ScandiBERT	87.9 $_{0.4}$	93.1 $_{0.3}$	83.2 $_{1.0}$
	mDeBERTa	86.6 $_{0.5}$	92.6 $_{0.2}$	81.2 $_{0.6}$

Table 9: **Test scores of baseline models on slot filling for F_1 variants: loose, unlabelled, and strict span** (all F_1 scores in %). Strict span is the F_1 score we use throughout, where both the span and label must be fully correct, loose F_1 allows for partial matches of the span (if the label is correct), and unlabelled ignores the label (considering only the span overlaps). The results are averaged over three runs, with standard deviations as subscripts.

		Intents				Slots			
PLM	**Split**	r	p_r	ρ	p_ρ	r	p_r	ρ	p_ρ
mDeBERTa	dev	–0.51	0.09	–0.60	0.04	–0.56	0.06	–0.45	0.14
	test	–0.38	0.22	–0.44	0.15	–0.42	0.17	–0.35	0.27
	dev+test	–0.36	0.08	–0.44	0.03	–0.47	0.02	–0.39	0.06
ScandiBERT	dev	–0.57	0.06	–0.56	0.06	–0.24	0.44	–0.34	0.28
	test	–0.66	0.02	–0.68	0.02	0.14	0.66	0.07	0.83
	dev+test	–0.53	0.01	–0.50	0.01	–0.16	0.45	–0.19	0.36
NorBERT	dev	–0.70	0.01	–0.63	0.03	–0.16	0.63	–0.30	0.34
	test	–0.82	0.00	–0.81	0.00	–0.03	0.93	–0.09	0.79
	dev+test	–0.49	0.01	–0.56	0.00	–0.18	0.40	–0.27	0.20

Table 10: **Correlations between the split word ratio difference and SID performance for the noising experiments:** Pearson's r and Spearman's ρ with corresponding p-values (p-values $\geq$0.05 have a grey background). Each dev or test row is based on twelve observations (four noise levels à three initializations).

Task		Intents Dev	Intents Δ_{dev}	Intents Test	Intents Δ_{test}	Slots Dev	Slots Δ_{dev}	Slots Test	Slots Δ_{test}	Aux. task performance (dev) Dial	POS	Dep	NER
English SID training data													
Baseline		96.4$_{0.5}$		94.8$_{0.8}$		81.3$_{0.3}$		80.7$_{0.7}$					
Dial	$\times$	84.3$_{2.7}$	–12.1	83.8$_{3.2}$	–11.0	75.8$_{1.8}$	–5.5	75.8$_{1.2}$	–4.9	75.9$_{3.5}$			
	$\rightarrow$	95.8$_{0.7}$	–0.6	94.0$_{1.7}$	–0.8	79.7$_{1.3}$	–1.6	79.2$_{0.9}$	–1.5	80.0$_{0.3}$			
POS	$\times$	96.0$_{0.5}$	–0.4	94.9$_{0.3}$	+0.0	81.6$_{0.2}$	+0.3	81.1$_{0.3}$	+0.4		79.5$_{0.6}$		
	$\rightarrow$	96.3$_{0.2}$	–0.1	94.7$_{0.2}$	–0.2	82.3$_{0.8}$	+1.0	82.2$_{1.1}$	+1.5		92.1$_{0.0}$		
Dep	$\times$	94.7$_{1.2}$	–1.8	93.5$_{0.6}$	–1.3	82.0$_{0.4}$	+0.7	81.5$_{0.2}$	+0.8			46.6$_{0.8}$	
	$\rightarrow$	96.7$_{1.0}$	+0.3	94.9$_{1.2}$	+0.1	82.5$_{0.9}$	+1.2	81.8$_{0.7}$	+1.1			67.8$_{0.6}$	
NER	$\times$	97.2$_{0.9}$	+0.8	95.3$_{1.0}$	+0.5	80.9$_{0.9}$	–0.4	80.6$_{1.0}$	–0.1				93.2$_{0.2}$
	$\rightarrow$	96.8$_{0.4}$	+0.4	95.0$_{0.4}$	+0.1	81.3$_{1.1}$	+0.0	81.1$_{0.9}$	+0.4				93.0$_{0.1}$
Machine-translated Norwegian SID training data													
Baseline		97.6$_{0.0}$		96.3$_{0.1}$		55.5$_{0.4}$		54.6$_{0.4}$					
Dial	$\times$	89.8$_{1.4}$	–7.8	89.2$_{1.4}$	–7.1	53.0$_{0.1}$	–2.6	51.7$_{0.1}$	–2.9	77.1$_{1.2}$			
	$\rightarrow$	96.1$_{1.0}$	–1.5	95.2$_{1.0}$	–1.1	54.4$_{0.4}$	–1.2	53.7$_{0.5}$	–0.9	79.7$_{0.3}$			
POS	$\times$	97.9$_{0.3}$	+0.3	96.8$_{0.4}$	+0.4	54.0$_{0.5}$	–1.5	53.7$_{0.6}$	–0.9		70.3$_{1.4}$		
	$\rightarrow$	97.8$_{0.5}$	+0.2	96.7$_{0.4}$	+0.3	55.5$_{0.4}$	+0.0	54.4$_{0.8}$	–0.2		92.1$_{0.1}$		
Dep	$\times$	98.0$_{0.4}$	+0.4	96.9$_{0.3}$	+0.5	54.5$_{0.2}$	–1.0	53.7$_{0.2}$	–0.9			41.1$_{0.9}$	
	$\rightarrow$	97.5$_{0.7}$	–0.1	96.4$_{0.3}$	+0.1	55.7$_{0.5}$	+0.2	54.8$_{0.6}$	+0.2			67.8$_{0.6}$	
NER	$\times$	97.9$_{0.6}$	+0.3	96.9$_{0.1}$	+0.6	54.6$_{0.5}$	–0.9	53.8$_{0.3}$	–0.8				92.3$_{0.3}$
	$\rightarrow$	97.6$_{0.2}$	–0.0	96.4$_{0.5}$	+0.1	54.1$_{0.6}$	–1.4	53.5$_{1.0}$	–1.1				93.0$_{0.1}$

Table 11: **Performance of the models trained on auxiliary task data** on the SID data (development and test) and the auxiliary tasks (development sets). Scores are averaged over three runs (standard deviations in subscript numbers) and in % – intent classification: accuracy, slot filling: span F_1, dialect classification (“dial”): accuracy, POS tagging: accuracy, dependency parsing (“dep”): labelled attachment score, NER: span F_1. The Δ columns show the differences to the respective baseline. Joint multi-task learning is denoted by a $\times$, and intermediate-task training by a $\rightarrow$.

Trained on auxiliary tasks and English SID data

	Intents (acc., %)										**Slots (span F_1, %)**									
Aux	**B**	Δ_B	**N**	Δ_N	**T**	Δ_T	**W**	Δ_W	**all**	Δ_{all}	**B**	Δ_B	**N**	Δ_N	**T**	Δ_T	**W**	Δ_W	**all**	Δ_{all}
none	96.3$_{0.7}$		93.3$_{1.5}$		94.0$_{0.9}$		95.6$_{0.5}$		94.8$_{0.8}$		83.7$_{1.0}$		75.7$_{0.8}$		79.3$_{0.7}$		82.8$_{0.8}$		80.7$_{0.7}$	
	9.2%		*18.5%*		*26.9%*		*45.4%*				*9.2%*		*18.5%*		*26.9%*		*45.4%*			
Dial ×	87.3$_{0.6}$	–9.1	72.8$_{6.1}$	–20.4	81.0$_{5.0}$	–13.0	89.2$_{1.7}$	–6.5	83.8$_{3.2}$	–11.0	82.7$_{0.9}$	–1.0	70.0$_{2.2}$	–5.7	70.1$_{1.8}$	–9.3	79.9$_{0.8}$	–2.9	75.8$_{1.2}$	–4.9
Dial →	96.5$_{0.9}$	+0.1	92.0$_{3.5}$	–1.3	92.0$_{1.6}$	–2.0	95.5$_{1.3}$	–0.2	94.0$_{1.7}$	–0.8	81.9$_{0.8}$	–1.9	75.4$_{1.8}$	–0.4	76.2$_{1.5}$	–3.2	82.0$_{0.2}$	–0.9	79.2$_{0.9}$	–1.5
	0.0%		*28.1%*		*7.6%*		*33.7%*				*0.0%*		*28.1%*		*7.6%*		*33.7%*			
POS ×	95.9$_{0.9}$	–0.5	93.1$_{0.6}$	–0.2	94.0$_{0.9}$	+0.0	95.9$_{0.4}$	+0.2	94.9$_{0.3}$	+0.0	83.7$_{0.5}$	–0.1	77.0$_{0.2}$	+1.3	80.5$_{0.8}$	+1.2	82.6$_{0.3}$	–0.2	81.1$_{0.3}$	+0.4
POS →	96.2$_{0.4}$	–0.1	92.8$_{0.6}$	–0.5	93.7$_{0.2}$	–0.4	95.7$_{0.5}$	+0.1	94.7$_{0.2}$	–0.2	85.1$_{0.7}$	+1.3	77.3$_{1.7}$	+1.6	81.5$_{1.2}$	+2.2	83.8$_{0.9}$	+1.0	82.2$_{1.1}$	+1.5
Dep ×	95.0$_{0.7}$	–1.3	91.6$_{0.8}$	–1.6	91.2$_{1.7}$	–2.8	95.4$_{0.1}$	–0.3	93.5$_{0.6}$	–1.3	83.8$_{0.6}$	+0.0	77.7$_{0.8}$	+2.0	80.4$_{0.3}$	+1.1	83.1$_{0.2}$	+0.3	81.5$_{0.2}$	+0.8
Dep →	95.9$_{1.2}$	–0.5	93.0$_{1.3}$	–0.2	94.3$_{1.9}$	+0.3	95.8$_{0.8}$	+0.1	94.9$_{1.2}$	+0.1	84.0$_{0.8}$	+0.3	77.5$_{1.9}$	+1.8	80.3$_{1.2}$	+0.9	83.9$_{0.4}$	+1.0	81.8$_{0.7}$	+1.1
	100.0%		*0.0%*		*0.0%*		*0.0%*				*100.0%*		*0.0%*		*0.0%*		*0.0%*			
NER ×	96.4$_{0.7}$	+0.1	94.1$_{1.4}$	+0.8	95.0$_{1.5}$	+0.9	95.8$_{0.6}$	+0.2	95.3$_{1.0}$	+0.5	83.9$_{2.2}$	+0.2	76.2$_{1.0}$	+0.5	79.1$_{1.0}$	–0.2	82.5$_{0.7}$	–0.4	80.6$_{1.0}$	–0.1
NER →	96.3$_{0.2}$	–0.1	93.6$_{0.9}$	+0.4	94.5$_{0.7}$	+0.5	95.5$_{0.2}$	–0.2	95.0$_{0.4}$	+0.1	84.5$_{2.0}$	+0.7	75.9$_{0.4}$	+0.2	79.3$_{1.6}$	+0.0	83.5$_{0.4}$	+0.7	81.1$_{0.9}$	+0.4

Trained on auxiliary tasks and machine-translated Norwegian SID data

	Intents (acc., %)										**Slots (span F_1, %)**									
Aux	**B**	Δ_B	**N**	Δ_N	**T**	Δ_T	**W**	Δ_W	**all**	Δ_{all}	**B**	Δ_B	**N**	Δ_N	**T**	Δ_T	**W**	Δ_W	**all**	Δ_{all}
none	97.4$_{0.0}$		94.9$_{0.1}$		96.9$_{0.5}$		96.3$_{0.2}$		96.3$_{0.1}$		58.7$_{0.3}$		50.9$_{1.1}$		54.6$_{0.9}$		55.2$_{0.4}$		54.6$_{0.4}$	
	9.2%		*18.5%*		*26.9%*		*45.4%*				*9.2%*		*18.5%*		*26.9%*		*45.4%*			
Dial ×	95.9$_{0.8}$	–1.5	80.1$_{4.1}$	–14.8	86.4$_{2.6}$	–10.5	93.3$_{0.6}$	–3.0	89.2$_{1.4}$	–7.1	57.3$_{0.8}$	–1.4	46.9$_{0.7}$	–4.1	50.3$_{0.5}$	–4.4	53.3$_{0.2}$	–1.9	51.7$_{0.1}$	–2.9
Dial →	97.6$_{0.0}$	+0.2	92.7$_{0.9}$	–2.3	95.5$_{0.9}$	–1.4	95.7$_{1.4}$	–0.6	95.2$_{1.0}$	–1.1	58.8$_{0.9}$	+0.1	51.0$_{0.2}$	+0.0	52.6$_{0.8}$	–2.0	54.4$_{0.7}$	–0.8	53.7$_{0.5}$	–0.9
	0.0%		*28.1%*		*7.6%*		*33.7%*				*0.0%*		*28.1%*		*7.6%*		*33.7%*			
POS ×	97.6$_{0.4}$	+0.2	95.6$_{0.6}$	+0.6	97.2$_{0.2}$	+0.3	96.8$_{0.5}$	+0.5	96.8$_{0.4}$	+0.4	57.7$_{0.7}$	–1.0	50.6$_{0.2}$	–0.4	53.9$_{1.3}$	–0.8	54.0$_{0.5}$	–1.2	53.7$_{0.6}$	–0.9
POS →	97.5$_{0.4}$	+0.1	95.2$_{0.7}$	+0.3	97.3$_{0.3}$	+0.4	96.7$_{0.4}$	+0.4	96.7$_{0.4}$	+0.3	58.2$_{0.8}$	–0.4	51.3$_{0.6}$	+0.4	54.2$_{0.6}$	–0.4	54.9$_{1.1}$	–0.3	54.4$_{0.8}$	–0.2
Dep ×	97.1$_{0.1}$	–0.3	95.8$_{0.6}$	+0.8	97.2$_{0.4}$	+0.4	97.0$_{0.1}$	+0.7	96.9$_{0.3}$	+0.5	57.9$_{0.6}$	–0.8	50.8$_{0.1}$	–0.2	53.9$_{0.1}$	–0.7	53.9$_{0.3}$	–1.3	53.7$_{0.2}$	–0.9
Dep →	97.5$_{0.3}$	+0.1	95.0$_{0.5}$	+0.1	97.1$_{0.2}$	+0.2	96.4$_{0.5}$	+0.1	96.4$_{0.3}$	+0.1	59.0$_{0.8}$	+0.3	52.0$_{0.4}$	+1.0	54.9$_{0.8}$	+0.3	55.0$_{0.7}$	–0.2	54.8$_{0.6}$	+0.2
	100.0%		*0.0%*		*0.0%*		*0.0%*				*100.0%*		*0.0%*		*0.0%*		*0.0%*			
NER ×	97.8$_{0.0}$	+0.4	95.4$_{0.4}$	+0.5	97.3$_{0.1}$	+0.4	97.0$_{0.0}$	+0.7	96.9$_{0.1}$	+0.6	58.5$_{0.2}$	–0.1	50.5$_{0.7}$	–0.5	54.0$_{0.6}$	–0.7	54.1$_{0.3}$	–1.1	53.8$_{0.3}$	–0.8
NER →	97.3$_{0.6}$	–0.1	95.1$_{0.6}$	+0.2	96.6$_{0.5}$	–0.3	96.6$_{0.5}$	+0.3	96.4$_{0.5}$	+0.1	57.7$_{1.2}$	–0.9	50.5$_{2.0}$	–0.5	52.6$_{1.1}$	–2.0	54.4$_{0.9}$	–0.8	53.5$_{1.0}$	–1.1

Table 12: **Dialect-wise test results of the models trained on auxiliary tasks.** The numbers in italics with blue backgrounds describe the dialect distributions in the data used to train the respective auxiliary tasks (e.g., 28.1% of the training data for the syntactic tasks is in North Norwegian). Key: B = Bokmål, N = North N., T = Trønder N., W = West Norwegian, Δ = difference to the baseline model (in pp.), $\times$ = multi-task learning, $\rightarrow$ = intermediate-task training.