LTG at VarDial 2025 NorSID: More and Better Training Data for Slot and Intent Detection

Marthe Midtgaard, Petter Mæhlum, Yves Scherrer

University of Oslo, Department of Informatics

{ marthemi | pettemae | yvessc } @ifi.uio.no

Abstract

This paper describes the LTG submission to the VarDial 2025 shared task, where we participate in the Norwegian slot and intent detection subtasks. The shared task focuses on Norwegian dialects, which present challenges due to their low-resource nature and variation. We test a variety of neural models and training data configurations, with the focus on improving and extending the available Norwegian training data. This includes automatically re-aligning slot spans in Norwegian Bokmål, as well as re-translating the original English training data into both Bokmål and Nynorsk. We also re-annotate an external Norwegian dataset to augment the training data. Our best models achieve first place in both subtasks, achieving an span F1 score of 0.893 for slot filling and an accuracy of 0.980 for intent detection. Our results indicate that while translation quality is less critical, improving the slot labels has a notable impact on slot performance. Moreover, adding more standard Norwegian data improves performance, but incorporating even small amounts of dialectal data leads to greater gains.

1 Introduction

The task of spoken language understanding (SLU) is an essential part of task-oriented dialogue systems and voice assistants like Siri and Alexa. SLU consists in annotating and identifying the meaning of spoken prompts, and typically comprise an Automatic Speech Recognition (ASR) component for converting audio to text, alongside a Natural Language Understanding (NLU) component for extracting the semantic meaning of the utterance (Faruqui and Hakkani-Tür, 2022).

Slot and intent detection (SID), also known as slot filling and intent classification, is a key task in NLU. Intent classification categorizes an entire user utterance into a predefined intent class, determining the purpose or goal behind the user's utterance. On the other hand, slot filling is a span labeling task that assigns each token in an utterance a label, capturing the essential information required to fulfill each intent, such as dates, locations and names. An example is shown in Figure 1.

Intent	Utterance
PlayMusic	play with or without you by U2

Figure 1: Example utterance annotated with slots and intent. Pink: track, green: artist.

While significant progress has been made in the field of NLU, the continued development of SID models relies on the availability of datasets annotated with slots and intents. In low-resource scenarios, where little to no labeled data is available, challenges emerge in developing accurate SID models. Over the years, there has been a notable increase in research on various low-resource scenarios, and VarDial has provided an important venue for discussion and research in handling linguistic diversity and low-resource scenarios (Aepli et al., 2023).

The 2025 iteration of the VarDial Shared Task (Scherrer et al., 2025) introduces the novel `NorSID` dataset to tackle the low-resource nature of Norwegian dialects. This dataset includes prompts intended for digital assistants across ten Norwegian dialects as well as Norwegian Bokmål. Each prompt is annotated with a dialect label, an intent label, and slot spans following the BIO scheme. `NorSID` therefore forms the foundation of this Shared Task, which includes three subtasks: dialect identification, intent detection, and slot filling. Our team participated in the latter two.

We make the following main contributions:

1. We compare various pre-trained models – both multilingual and Norwegian-specific ones – and fine-tune them on the $\texttt{xSID}_{0.6}$ (van der Goot et al., 2021a; Aepli et al., 2023; Winkler et al., 2024) data in English, Danish and Norwegian.

Proceedings of the Twelfth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 200–208
January 19, 2025.

2. To enhance the quality of the Norwegian training data, we create a re-aligned version as well as re-translations from English into both Bokmål and Nynorsk.

3. We use the existing Norwegian split of the MASSIVE dataset (FitzGerald et al., 2023) and convert its annotations to the xSID$_{0.6}$ annotation scheme.[1]

2 Data and Evaluation

For developing our Norwegian SID models, the xSID$_{0.6}$ and NorSID datasets serve as our foundational resources. To address the challenge of limited annotated Norwegian data, we experiment with utilizing parts of the Norwegian split of MASSIVE to augment the training data.

xSID$_{0.6}$ is a recent NLU dataset, serving as a benchmark for cross-lingual transfer with data in 17 languages, including 5 low-resource languages and dialects. Although Norwegian is not part of xSID$_{0.6}$, a projected training set was created specifically for this Shared Task by translating the English training data into Norwegian and aligning the slots in the same way as for the non-English xSID$_{0.6}$ training data.

xSID$_{0.6}$ is derived from the English NLU datasets Facebook (Schuster et al., 2019) and Snips (Coucke et al., 2018), where the original English development and test data were manually translated and re-annotated into the other languages. For high-resource languages, the training data was machine-translated and slots were aligned using attention (van der Goot et al., 2021a). The final xSID$_{0.6}$ dataset is annotated with 18 intents and 41 slots. It includes 43.6k training utterances for high-resource languages, along with 300 development and 500 test utterances for all languages.

NorSID is based on the NoMusic corpus (Mæhlum and Scherrer, 2024), and is a Norwegian extension of xSID$_{0.6}$, with parallel data for 10 Norwegian dialects along with Bokmål (B). The dialects are grouped into 3 dialect areas, West Norwegian (V), North Norwegian (N) and Trøndersk (T). The dataset consists of translations of the validation and test splits from xSID$_{0.6}$, annotated with the same slots and intents. Each utterance is translated into all dialects by native speakers who use dialectal writing on a regular basis, and slots are manually annotated by native NLP professionals. By including several renditions of semantically identical utterances, dialectal diversity is showcased, e.g., as indicated by lexical and syntactic differences, and this diversity introduces novel opportunities to enhance the robustness of both training and evaluation of Norwegian SLU systems.

MASSIVE stands as the largest multilingual SLU dataset to date, with "1M realistic, human-created, labeled virtual assistant utterances" (FitzGerald et al., 2023). The dataset comprises 51 languages, with 19.5k utterances per language over 18 domains, 60 intents and 55 slots. MASSIVE is thus more comprehensive than xSID$_{0.6}$ and, with some overlapping slots and intents, serves as a suitable resource for augmenting xSID$_{0.6}$.

MASSIVE is, to our knowledge, the only other SID annotated dataset that includes Norwegian, but it only contains utterances in Norwegian Bokmål, which limits its ability to capture the diverse nature of the Norwegian language. With two official written standards, Bokmål and Nynorsk, as well as numerous dialectal variations, the effect of MASSIVE might be limiting on the dialects.

Although limited to Bokmål, MASSIVE still offers a valuable resource worth exploring. The process of aligning Norwegian MASSIVE utterances to the xSID$_{0.6}$ scheme is described in section 4.3. However, since other SID annotated datasets are not allowed in the Shared Task, we provide the models based on MASSIVE outside of the competition.

Evaluation The evaluation of the slot and intent subtasks is based on two primary metrics: span F1 score for slot filling and accuracy for intent detection. For span F1, both the span and slot label must match the gold standard for the prediction to be counted as correct. Intent accuracy measures the proportion of correct intent predictions out of the total number of utterances. Additionally, the models will be evaluated using dialect-specific slot and intent scores to assess robustness to dialectal variation. We use the official evaluation script provided by the organizers.

3 Existing Data and Pre-Trained Models

In recent years, jointly addressing the tasks of slot and intent detection has been recognized as an effective strategy (Weld et al., 2022). In this work, we adopt this joint approach by utilizing the MaChAmp$_{0.4.2}$ toolkit (van der Goot et al., 2021b)

[1] Our contributions are available at: https://github.com/marthemidtgaard/SID-for-Norwegian-dialects

with default hyperparameters for all experiments. We evaluate the models on the NorSID development set, and compare results to the results of mBERT (Devlin et al., 2019), following the setup in van der Goot et al. (2021a).

3.1 Pre-Trained Models

As a first experiment, we investigate which pre-trained base models are the most suitable for the task. We use several multilingual models – XLM-R-large (Conneau et al., 2020), RemBERT (Chung et al., 2020), mT0-base (Muennighoff et al., 2023), mDeBERTaV3-base (He et al., 2022) – all of which include Norwegian Bokmål in their training data and have demonstrated state-of-the-art performance on zero-shot cross-lingual tasks. These models are therefore expected to exhibit enhanced robustness in the low-resource scenario of Norwegian (Artemova et al., 2024). Additionally, we explore two Norwegian-specific models – NB-BERT-base[2] and NorBERT-base-3 (Samuel et al., 2023) – which are trained on both Bokmål and Nynorsk. These models may therefore offer improved performance when applied to the SID task for Norwegian dialects.

3.2 Fine-Tuning Languages

The $\text{xSID}_{0.6}$ training data is available in various languages, but all except the English data was machine-translated, with potentially poor translation quality. We identify three languages for our next experiments: English, Danish and Norwegian. We consider these languages to be the most effective, as they exhibit the closest linguistic proximity to Norwegian dialects among the languages in $\text{xSID}_{0.6}$. We fine-tune three separate models for each pre-trained model to understand how annotation quality and linguistic proximity affect the prediction performance.

3.3 Results

The results of these experiments on the development set are presented in Table 1. They reveal notable trends in performance across different models. For intent classification, fine-tuning on Norwegian data yields the best accuracy for almost all models. For slot filling, the opposite trend is observed: performance drops considerably when models are fine-tuned on Norwegian or Danish data. In this case, fine-tuning on English achieves much better results, which can be attributed to the fact that the original training data is in English. Norwegian and Danish training data, derived through machine translation and slot alignment via attention mechanisms, likely suffer from noise and alignment inconsistencies, which impacts the performance.

Model	Slots			Intents		
	en	da	nb	en	da	nb
mBERT	.659	.525	.569	.898	.907	.907
XLM-R	.800	.566	.561	.985	.984	.979
RemBERT	.734	.558	.549	.944	.962	.974
mT0	.737	.531	.529	.889	.922	.921
mDeBERTa	.787	.579	.564	.965	.934	.982
NB-BERT	**.812**	.590	.572	.988	.969	.989
NorBERT	.797	.568	.558	.964	.964	**.990**

Table 1: Results on slot filling (F1) and intent detection (accuracy) on the dev set. **Bold:** Top intent accuracy and span F1 score.

In terms of base models, NB-BERT delivers the overall best results across subtasks and languages, followed by XLM-R and NorBERT. All three models outperform the baseline mBERT. NorBERT fine-tuned on Norwegian has the highest intent accuracy, but NB-BERT achieves the highest increase in intent accuracy compared to the baseline mBERT, with a +0.90 improvement when fine-tuned on English. For slot filling, NB-BERT sees smaller gains compared to the mBERT baseline: +0.153 when fine-tuned on English and only +0.003 when fine-tuned on Norwegian. These results highlight the difficulty of slot filling and the need for further refinement to improve performance.

Based on these findings, our continued experiments focus on the top-performing models NB-BERT, XLM-R and NorBERT. Fine-tuning on English emerges as the best approach for slot filling, while fine-tuning on Norwegian is most effective for intent classification. We also conduct additional experiments combining English and Norwegian training data, aiming to benefit from their complementary strengths.

4 Improving and Extending the Training Data

To address the lower span F1 scores observed when fine-tuning on Norwegian data, we further explore ways to improving the quality of the training data.

[2] https://github.com/NbAiLab/notram

Dataset	B-tags	I-tags	Sum
en	82,408	61,036	143,444
nb	82,644	91,556	174,200
nb_ra	87,411	45,340	132,760
nb_rt	83,165	47,143	130,308
nn_rt	84,644	45,563	130,207

Table 2: Distributions of B- and I-tags. The number of B-tags corresponds to the number of slot spans.

4.1 Slot Re-Alignment

A comparison of slot counts shows that 143,444 English tokens are annotated with a slot, compared to 174,200 in Norwegian (nb), despite Norwegian having slightly fewer tokens (336,387 vs 341,094). This suggests an overuse of slots in Norwegian, driven by Norwegian having around 30k more I-tags. The distribution of B- and I-tags is shown in Table 2. For example, `I-datetime` is used 14,153 more times in Norwegian. This indicates poor slot projection quality and highlights the need for re-alignment to improve training effectiveness.

To project labels from English to Norwegian, we use `simAlign` (Jalili Sabet et al., 2020), a word alignment tool that leverages both static and contextualized embeddings to map English tokens to their Norwegian counterparts. Challenges arise when multiple English tokens align with a single Norwegian token, as each Norwegian token can hold only one slot. If all aligned English tokens share the same slot, it is transferred directly. This is typically the case for compound words, which are split across multiple tokens in English but generally appear as a single token in Norwegian. For example for *rain forecast*, the slot `weather/attribute` is easily transferred to the Norwegian *regnvarsel*. For conflicting slots, we calculate cosine similarity between the contextualized embeddings of each English token and the Norwegian token using `XLM-R`, considering a context window of two tokens before and after each token. The English token with the highest similarity score is selected, and its slot is transferred to the Norwegian token. After alignment, we reapply the BIO tagging format and adjust slot spans based on the $\texttt{xSID}_{0.6}$ annotation guidelines (van der Goot et al., 2021a), excluding prepositions like *på*, *for*, and *til*, and the infinitive marker *å* from the edges of slot spans.

This results is a new Norwegian training set (`nb_ra`), where `ra` stands for re-alignment. The updated set contains 132,760 slots – 41,440 fewer than the original `nb` version – bringing it closer to the total number of slots in the English dataset (see Table 2). The slot spans in `nb_ra` are generally shorter, with about 50% fewer I-tags compared to `nb`. This reduction arises primarily because fewer surrounding tokens are included in slot spans. The example in Figure 2 illustrates this difference.

en	will it rain today ?
nb	Kommer det til å regne i dag ?
nb_ra	Kommer det til å regne i dag ?

Figure 2: Examples of slots. Green: weather/attribute, pink: datetime.

4.2 Re-Translation

Manual inspection of the original Norwegian translations reveals significant translation issues. For example, the original translation model may mistranslate questions into declaratives or with atypical word order (see example 1 in Figure 3). The original translation also suffers from unknown tokens, such as *february* being translated as *<unk> ary*. In addition, the translation model often splits expression into multiple tokens due to punctuations, leading to misaligned tokens and incorrect slot transfers in `nb` and `nb_ra`. This is for example frequent in time expressions as in example 2 of Figure 3. Improving the quality of the translations therefore seems essential to enhance slot alignment and further increase span F1 scores.

We re-translated the English $\texttt{xSID}_{0.6}$ training data into Bokmål and Nynorsk by using `NorMistral-7b-warm`,[3] which is an LLM initialized from `Mistral-7B-v0.1`[4], and continuously pre-trained on Norwegian data. `NorMistral-7b-warm` was chosen for its favorable performance in prior zero-shot English-to-Bokmål and English-to-Nynorsk translation evaluations.[3]

The original data contains inconsistent use of proper capitalization and punctuation. The first part is problematic since the dataset contains numerous proper names that should not be translated into Norwegian, and to improve the quality of the translations, we apply truecasing to each sentence

[3] https://huggingface.co/norallm/normistral-7b-warm

[4] https://huggingface.co/mistralai/Mistral-7B-v0.1

en	is Tuesday to be rainy
nb_ra	Det regner på torsdager
	(It rains on Thursdays)
nb_rt	Skal det bli regn på tirsdag ?
	(Will it be rain on Tuesday ?)
nn_rt	Kjem det til å bli regn på tysdag
	(Will it come to be rain on Tuesday ?)
en	Set alarm for 5:30 am tomorrow
nb_ra	Alarm kl . 17 . 30 i morgen .
	(Alarm at 17 : 30 in the morning)
nb_rt	Sett alarm til kl. 05.30 i morgen
	(Set alarm to 5:30 am tomorrow)
nn_rt	Set alarm for 5.30 i morgon
	(Set alarm for 5:30 am tomorrow)

Figure 3: Examples of slots. Green: weather/attribute, pink: datetime.

using the Python `truecase`[5] library. This results in two new Norwegian datasets, `nb_rt` (Bokmål re-translated) and `nn_rt` (Nynorsk re-translated), which both undergo the same slot alignment as `nb_ra`. Our decision to include Nynorsk is motivated by the fact that it more closely resembles many Norwegian dialects than Bokmål. This makes Nynorsk potentially more valuable for capturing linguistic features representative of dialectal variation.

Manual inspection of the new translations reveals significant improvement over `nb`, both in the choice of words and sentence structure. The new model produces structurally accurate sentences better reflecting the English source (see example 1 in Figure 3). The issue of unknown tokens is also entirely resolved in the new dataset.

Since the Norwegian re-translations follow the same alignment process as `nb_ra`, some of the same alignment issues remain. For example, syntactic differences between English and Norwegian can challenge the alignment and map tokens based on their position in the sentence (see example 1 in Figure 3). However, the improved translations reduce unnecessary token splitting, particularly for time expressions, resulting in better slot labeling.

[5] https://github.com/daltonfury42/truecase

4.3 Adapting the Norwegian MASSIVE Dataset

As another means to improve span F1 scores, we follow the approach of Winkler et al. (2024), who propose to extract utterances from the `MASSIVE` dataset that align with intents in $\texttt{xSID}_{\texttt{0.6}}$ and to re-annotating them following the $\texttt{xSID}_{\texttt{0.6}}$ annotation guidelines. While `MASSIVE` contains a broader range of intents, Winkler et al. (2024) successfully identified 2021 utterances matching the $\texttt{xSID}_{\texttt{0.6}}$ intents. The mapping and re-annotation process is documented in Appendix B of their work (Winkler et al., 2024).

Building on their efforts, we use their mapped Bavarian utterances to identify the corresponding Norwegian utterances in `MASSIVE`. Intents were directly transferred from the Bavarian dataset, while slots had to be manually annotated.

Although we aimed to follow the slots of Winkler et al. (2024), we found deviations from the slot-intent combinations in $\texttt{xSID}_{\texttt{0.6}}$. For example, they apply the `object_select` slot to several tokens, whereas $\texttt{xSID}_{\texttt{0.6}}$ restricts this slot to the `RateBook` intent, leaving similar tokens in other intents unannotated. In such cases, we diverged from the choices of Winkler et al. (2024) and adhered strictly to the slot–intent combinations in $\texttt{xSID}_{\texttt{0.6}}$, ensuring that the model learns patterns consistent with those in $\texttt{xSID}_{\texttt{0.6}}$. This results in a new Norwegian Bokmål training dataset named `nb_mas`.[6]

4.4 Results

Re-alignment The fine-tuning results on the `NorSID` development set using our best-performing models on `nb_ra` are presented in Table 3. For slot filling, the `nb_ra` dataset shows substantial improvements over `nb` across all models. For example, the F1 score for `NB-BERT` increases from 0.575 (`nb`) to 0.762 (`nb_ra`), a gain of nearly 33%. Similar improvements are observed for `XLM-R` and `NorBERT`, and these enhancements indicate that the re-alignment process helps improve slot annotations. In addition, adding English data to `nb_ra` (`en+nb_ra`) further boosts performance for `NB-BERT` and `NorBERT`, as it allows the models to leverage the higher-quality English slot annotations. The linguistic similarities between English and Norwegian slots, such as named entities, enable the models to learn transferable cross-lingual patterns,

[6] The re-annotated and re-translated data, as well as the Norwegian MASSIVE data, are available at: https://github.com/marthemidtgaard/SID-for-Norwegian-dialects.

	Slots			Intents		
	NB-B.	XLM-R	NorB.	NB-B.	XLM-R	NorB.
en	**.812**	.800	.797	.988	.985	.964
nb	.572	.561	.558	.989	.979	.990
nb_ra	.762	.764	.741	.987	.988	.986
en+nb_ra	.789	.761	.770	.994	.986	.993
nb_rt	.758	.751	.716	.987	.985	.984
en+nb_rt	.770	.761	.762	.991	.986	**.995**
nn_rt	.753	.753	.752	.981	.980	.986
en+nn_rt	.772	.783	.776	.992	.992	.982

Table 3: Results on slot filling (F1) and intent detection (accuracy) on the dev set. **Bold:** Top intent accuracy and span F1 score. nb_ra: re-aligned nb. nb_rt and nn_rt: machine translated and re-aligned nb/nn.

and the improved F1 score reflects the benefit of learning from the more accurate English data. However, the best performing new system, NB-BERT fine-tuned on en+nb_ra, still does not outperform fine-tuning solely on English. This suggests that the Norwegian data cannot match the quality of the English slot annotations, and that the inclusion of English only partially helps stabilize the noisier Norwegian annotations.

NB-BERT fine-tuned on en+nb_ra also achieves the highest intent accuracy (0.994), though the improvements from nb_ra are slightly smaller for intent detection compared to slot filling. This indicates that the inclusion of English data enhances performance without compromising the model's ability to understand Norwegian intents.

Re-translation Results from fine-tuning on the higher-quality translations, nb_rt and nn_rt, can be found in the bottom rows of Table 3. The en+nn_rt dataset achieves the highest F1 score (0.783 with XLM-R), with both XLM-R and NorBERT outperforming their en+nb_ra counterparts, and all three models surpassing their en+nb_rt counterparts. This likely reflects a closer resemblance between Nynorsk and Norwegian dialects compared to Bokmål, allowing the model to generalize better across dialectal variations. However, models trained exclusively on either nb_rt or nn_rt still fall significantly behind those trained on English as well, emphasizing the continued impact of higher-quality annotations in English.

Furthermore, the improved Bokmål translations in nb_rt show no notable impact on span F1 scores. Since the main changes from nb_ra are structural, slot alignments do not differ too much, resulting in comparable performance. Intent accuracy also remains stable across the new models, as our data augmentation efforts primarily target slot quality.

Overall, the findings underscore the need for further refinement in addressing slot alignment issues in order to bridge the performance gap between Norwegian and English. This is evident from the superior span F1 scores achieved by NB-BERT trained solely on English, which remains the best-performing model for slot labeling.

	Slots			Intents		
	NB-B.	XLM-R	NorB.	NB-B.	XLM-R	NorB.
en	.812	.800	.797	.988	.985	.964
+nb_mas	**.859**	.858	.832	.990	.985	.982
en+nb_ra	.789	.761	.770	**.994**	.986	.993
+nb_mas	.793	.799	.788	**.994**	.988	.992

Table 4: Impact of including the Norwegian MASSIVE (+nb_mas on slot filling (F1) and intent detection (accuracy), measured on the dev set. **Bold:** Top intent accuracy and span F1 score. nb_ra: re-aligned nb.

MASSIVE Fine-tuning results on the Norwegian MASSIVE data are shown in Table 4. Including nb_mas results in noticeable improvements in span F1 scores, with NB-BERT fine-tuned on en+nb_mas achieving the highest F1 score of 0.859, outperforming all other setups. This highlights the significant impact of MASSIVE data on slot performance when combined with high-quality English annotations. The other models also show notable improvements compared to their counterparts without MASSIVE.

Intent accuracy remains unaffected, suggesting that intent detection does not benefit from additional data. This is likely because the new utterances closely resemble those already present in $\text{xSID}_{0.6}$, indicating that they do not introduce novel patterns for the model to learn. This just shows that the intent mapping efforts by Winkler et al. (2024) were robust and effective.

Overall, these findings highlight the potential of including MASSIVE utterances to enhance slot filling. However, these models fall outside the permitted training data rules of the Shared Task and were submitted outside of the competition. Despite this, the promising results justify their inclusion in this paper to underscore the approach's potential.

5 Our Shared Task Submission

For our submission, we selected the best-performing model and training data combination per subtask. For slot filling, our best-performing

model is NB-BERT fine-tuned on English, while for intent detection, it is NorBERT fine-tuned on en+nb_rt. However, due to technical difficulties in test set prediction with NorBERT, we submitted our second-best intent detection model, namely NB-BERT fine-tuned on en+nb_ra, whose performance is nearly identical.

The Shared Task guidelines allowed the participants to use the development set for training. In order to further enhance the models mentioned above, we fine-tune them with the inclusion of the NorSID dev set. Since this prevents us from using a validation set, we submitted models after 20 epochs and after the best epoch. This resulted in three systems per subtask.

5.1 Results

Table 5 presents the official slot filling results, while Table 6 shows intent detection accuracy. Our models strongly outperform the mBERT baseline across both tasks, achieving a 38.6% improvement in slot filling with the top performing model fine-tuned on en+norsid. This model also outperforms the one fine-tuned on English, and the improved performance likely results from the close alignment between the dev and test sets, both translated by the same native speakers and annotated by the same team. By including the dev set in fine-tuning, utterances with the same style, word choices and slot spans are seen during training, facilitating improved performance on the test set, which closely resembles the training data.

Furthermore, Bokmål (B) consistently achieves the highest F1 scores, while North Norwegian (N) poses the greatest challenge, likely due to a greater linguistic divergence from the Bokmål and Nynorsk patterns learned during pre-training. This might also explain why North Norwegian, along with Trøndersk (T), sees the largest F1 improvements with the inclusion of the NorSID dev set, highlighting the importance of dialect-specific data in adapting the model to dialectal variations.

For intent detection, accuracy remains consistent across datasets and dialects, with North Norwegian performing only slightly lower than the others. This consistency suggests that intent detection effectively generalizes well across dialects and does not benefit from the inclusion of dialect utterances.

Interestingly, the number of training epochs has no impact on performance, suggesting rapid convergence due to the high quality data. Despite its small size, the NorSID development set provides sufficient task-specific information for effective optimization, with additional epochs offering no further gains or risk of overfitting.

ID	System	B	N	T	V	Overall
Baseline	mBERT	.715	.607	.632	.651	.644
LTG 1	en	.847	.801	.810	.833	.822
LTG 3	en+norsid (11)	.909	.872	.897	.895	.893
LTG 2	en+norsid (20)	.899	.879	.893	.896	.893
LTG 4	en+nb_mas+norsid	.918	.876	.890	.898	**.894**

Table 5: Dialect-specific and overall span F1 scores on the test set using NB-BERT. B=Bokmål, N=North Norwegian, T=Trøndersk, V=West Norwegian. Number of fine-tuning epochs in parentheses.

ID	System	B	N	T	V	Overall
Baseline	mBERT	.864	.826	.833	.848	.842
LTG 3	en+nb_ra+norsid (5)	.980	.972	.983	.982	**.980**
LTG 1	en+nb_ra	.982	.972	.983	.978	.979
LTG 2	en+nb_ra+norsid (20)	.982	.973	.981	.978	.979
LTG 4	en+nb_mas+norsid	.978	.967	.977	.972	.973

Table 6: Dialect-specific and overall intent accuracies on the test set using NB-BERT. Number of fine-tuning epochs in parentheses.

Finally, we also evaluate our best model including MASSIVE (en+nb_mas) on the test set and get a slot filling F1 score of 0.858. Dialect F1 scores, except for Bokmål, decrease significantly compared to en+norsid, indicating that new and unseen Bokmål utterances from MASSIVE contributes less than dialect utterances. This is also highlighted by the fact that dialect F1 scores are the same for en+norsid and en+nb_mas+norsid. However, interestingly, F1 score for Bokmål reaches a high with 0.918 for en+nb_mas+norsid. To further improve slot filling, a promising approach could be to add raw dialect data to fine-tuning, allowing the model to better handle nuanced dialect features.

6 Conclusion

In this paper, we presented our contribution to the two subtasks of the VarDial 2025 Shared Task: intent detection and slot filling. We evaluated different pre-trained models, including NB-BERT, XLM-R, and NorBERT, and identified NB-BERT as the best overall model, likely due to its superior ability to handle the linguistic complexities of Norwegian language varieties.

Slot filling emerged as a more challenging task than intent detection, with the latter showing consistent accuracy across experiments. This consistency can be attributed to the ease of transferring intents

from the original English xSID$_{0.6}$ to our different Norwegian versions, unlike slots, which rely on an automatic alignment process prone to errors. Our efforts to enhance slot annotations did not achieve the same level of performance as fine-tuning exclusively on English data, highlighting the critical role of high-quality slot annotations and the necessity for further refinement.

In addition, intent detection operates at the sentence level, relying on broader semantic features rather than the token-level distinctions critical for slot filling. As a result, it is less sensitive to dialectal variation and does not require extensive dialect-specific data. Models fine-tuned solely on Bokmål performed comparably to those incorporating dialectal data for intent detection. In contrast, slot filling is highly dependent on dialect-specific data due to token-level linguistic intricacies. For dialects, adding dialect-specific data proved more impactful than merely increasing the amount of Bokmål data.

Looking ahead, we aim to experiment with the inclusion of raw dialect data to better capture linguistic variation at the token level. Additionally, we intend to explore alternative methods for aligning slots between English and Norwegian to further enhance the quality of slot annotations.

Limitations

All of our models are trained once with a fixed random seed. This makes it hard to judge how stable the observed result patterns are. In particular for the intent detection task, many score differences are so small that they are likely due to random variation rather than to different training setups.

References

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.

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.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking Embedding Coupling in Pre-trained Language Models.

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. *arXiv preprint*. ArXiv:1805.10190 [cs].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Manaal Faruqui and Dilek Hakkani-Tür. 2022. Revisiting the Boundary between ASR and NLU in the Age of Conversational Dialog Systems. *Computational Linguistics*, 48(1):221–232. Place: Cambridge, MA Publisher: MIT Press.

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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual Generalization through Multitask Finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. 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.

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.

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.

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.

Henry Weld, Xiaoqi Huang, Siqu Long, Josiah Poon, and Soyeon Caren Han. 2022. A Survey of Joint Intent Detection and Slot Filling Models in Natural Language Understanding. *ACM Comput. Surv.*, 55(8):156:1–156:38.

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.