Zero-Shot Slot and Intent Detection in Low-Resource Languages

Sang Yun Kwon[1,⋆] **Gagan Bhatia**[1,⋆] **El Moatez Billah Nagoudi**[1]
Alcides Alcoba Inciarte[1] **Muhammad Abdul-Mageed**[1,2]
[1]Deep Learning & Natural Language Processing Group, The University of British Columbia
[2]Department of Natural Language Processing & Department of Machine Learning, MBZUAI
{skwon01@student.,gagan30@student.,moatez.nagoudi@,alcobaaj@mail.,muhammad.mageed@}ubc.ca

Abstract

Intent detection and slot filling are critical tasks in spoken and natural language understanding for task-oriented dialog systems. In this work we describe our participation in the slot and intent detection for low-resource language varieties (SID4LR; Aepli et al. (2023)). We investigate the slot and intent detection (SID) tasks using a wide range of models and settings. Given the recent success of multitask-prompted finetuning of large language models, we also test the generalization capability of the recent encoder-decoder model mT0 (Muennighoff et al., 2022) on new tasks (i.e., SID) in languages they have never intentionally seen. We show that our best model outperforms the baseline by a large margin (up to +30 F_1 points) in both SID tasks.

Lang.	Annotation
EN	Set an alarm for 6 am on Wed
GSW	Du em Mittwuch e Wecker dry fürem sächsi em Morge .
ST	Stell an Wecker firn Mittig af 6 in der friah
NAP	Imposta 'na sveglia 'e 6 'e matina 'e miercurì

Table 1: Examples of xSID annotations in our target languages from the validation set with intents (alarm / set_alarm) and slots (location , datetime). **EN**: English, **GSW**:Swiss German **ST**: South Tyrolean, **NAP**: Neapolitan

1 Introduction

Digital conversational assistants have become increasingly pervasive. Examples of popular virtual assistants include Siri, Alexa, and Google. A crucial factor in the effectiveness of these systems is their capacity to understand user input and respond or act accordingly to fulfill particular requirements. Most of these applications are voice-based and hence need spoken language understanding (SLU). SLU typically starts with automatic speech recognition (ASR), taking the sound of spoken language and transcribing it into text. Then, it handles natural language understanding (NLU) tasks to extract semantic features from the text including question answering, dialogue management, intent detection, and slot filling.

The intent detection task aims to recognize the speaker's desired outcome from a given utterance. And slot filling focuses on identifying the main arguments or the spans of words in the utterance that contain semantic information relevant to the intent. Table 1 shows four utterances in different languages: English, Swiss German (GSW), South Tyrolean (ST), and Neapolitan (NAP). The English example has the intent *set_alarm* and two individual spans *Set an alarm* and *6 am on Wed* are labeled with their slot tags *location* and *datetime*, respectively, using the Inside, Outside, Beginning (IOB) (Ramshaw and Marcus, 1995) tagging format.

In this work, we present our participation in the slot and intent detection for low-resource language varieties (SID4LR; Aepli et al. (2023)) shared task. The shared task takes as its target three low resources languages– Swiss German (GSW), South Tyrolean (ST), and Neapolitan (NAP). The main objective of the SID4LR shared task is to find the most effective approach for transferring knowledge to less commonly spoken languages that have limited resources and lack a standard writing system, in the zero-shot setting (i.e., without use of any training data). In the context of the shared task, we target the following four main research questions:

Q1: Can successful models on English SID tasks be generalizable to new unseen languages (i.e., the zero-shot setting)?

Q2: How do models trained on a language from the given language family fare on a low-resource variety from the same family under the zero-shot setting (i.e., with no access to training data from these low-resource varieties). For

Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), pages 241–250
May 5, 2023 ©2023 Association for Computational Linguistics

example, in our case, we ask how do models trained on German perform on Swiss German or South Tyrolean, and how do models trained on Italian perform on Neapolitan.

Q3: What impact does exploiting data augmentation techniques such as paraphrasing and machine translation have on the SID tasks in the zero-shot context?

Q4: Are the existing large multilingual models, trained using multitask-prompted fine-tuning, able to achieve zero-shot generalization to SID tasks in languages that they have never intentionally seen?

The rest of this paper is organized as follows: Section 2 is a literature review on intent and slot detection tasks. The shared task, the source data provided in SID4LR, and the external parallel data we exploit to build our models are described in Section 3. In Section 4, we provide information about datasets, baselines, and data preprocessing. The baseline, and multilingual pre-trained language models we used are described in Section 5. We present our experimental settings and our training procedures in Section 6. Section 7 is an analysis and discussion of our results. And we conclude in Section 8.

2 Related Work

The problem of low-resource slot and intent detection for languages with limited training data has been the focus of several recent research works. In this section, we discuss some of the most relevant and recent works, including datasets, benchmarks, and models that aim to address this challenge.

2.1 SID Benchmarks and Corpus

The table below provides an overview of various datasets used for NLU tasks. These datasets cover a range of languages, domains, intents, and slots, and are widely used to evaluate the performance of NLU models. Some of the prominent datasets include MASSIVE, SLURP, NLU Evaluation Data, ATIS, MultiATIS++, Snips, TOP, MTOP, Cross-lingual Multilingual Task-Oriented Dialog, Microsoft Dialog Challenge, and Fluent Speech Commands. These datasets have been used for tasks such as intent classification, slot filling, and semantic parsing. Overall, these datasets provide a useful resource for researchers to benchmark their models and develop better NLU systems.

2.2 SID Approaches and Models

The are many works devoted to the SID tasks. Most of these works are categorized into three approaches: *(1) single model for intent detection, (2) single model for slot filling, and (3) joint model.*

(1) Single Model for Intent Detection refers to developing a single model that can identify the intent behind a user's spoken or written input. This approach involves training a neural network or other machine learning model on a large dataset of labeled examples. Each example consists of user input and its corresponding intent label. The model then uses this training data to learn patterns and features that can accurately predict the intent of new user inputs. For instance, Ravuri and Stolcke (2015) proposed a recurrent neural network and LSTM models for intent detection in spoken language understanding. In this work, the authors first discuss the limitations of traditional intent detection approaches that rely on handcrafted features and propose using deep learning models to learn features directly from the data. Zhang et al. (2021) investigate the robustness of pre-trained transformers-based models such as BERT and RoBERTa for intent classification in spoken language understanding. They conduct experiments on two datasets, ATIS (Upadhyay et al., 2018) and SNIPS (Coucke et al., 2018), showing that pre-trained transformers perform well on in-scope intent detection.

(2) Single Model for Slot Filling is an approach that aims to develop a single model capable of identifying slots in spoken language understanding. The model takes a sentence as input and predicts the slot labels for each word in the sentence. Various recurrent neural network (RNN) architectures such as Elman-type (Mesnil et al., 2015) and Jordan-type (Mesnil et al., 2015) networks and their variants have been explored to find the most effective architecture for slot filling. Incorporating word embeddings has also been studied and found to improve slot-filling performance significantly. For example, Yao et al. (2014) use LSTM networks with word embeddings for slot filling on the ATIS (Upadhyay et al., 2018) dataset and achieve state-of-the-art (SOTA) results at the time. Goo et al. (2018) propose a bi-directional LSTM (BLSTM) with an attention mechanism for slot filling on the ATIS (Upadhyay et al., 2018) and SNIPS (Coucke et al., 2018) datasets.

(3) Joint Model is an approach that aims to jointly

Name	# Langs	Utt. per Lang (K)	Domains	Intents	Slots
Airline Travel Information System (ATIS) (Price, 1990)	1	5.8	1	26	129
ATIS with Hindi and Turkish (Upadhyay et al., 2018)	3	1.3-5.8	1	26	129
Cross-lingual Multilingual Task Oriented Dialog (Schuster et al., 2019)	3	5.08-43.3	3	12	11
Fluent Speech Commands (FSC) (Lugosch et al., 2019)	1	30	-	31	-
MASSIVE (FitzGerald et al., 2022)	51	19.5	18	60	55
Microsoft Dialog Challenge (Li et al., 2018)	1	38.2	3	11	29
MultiATIS++ (Xu et al., 2020)	9	1.4-5.8	1	21-26	99-140
Multilingual Task-Oriented Semantic Parsing (MTOP) (Li et al., 2021)	6	15.1-22.2	11	104-113	72-75
NLU Evaluation Data (Liu et al., 2019)	1	$25,7$	18	54	56
SLURP (Bastianelli et al., 2020)	1	$16,5$	18	60	55
SNIPS (Coucke et al., 2018)	1	14.4	-	7	53
Task Oriented Parsing (TOP) (Gupta et al., 2018)	1	44.8	2	25	36
xSID (van der Goot et al., 2021)	13	10	7	16	33

Table 2: SID benchmark and datasets with the number of languages covered, number of utterances per language, domain, intent count, and slot count.

model the intent detection and slot-filling tasks in spoken language understanding. This approach trains a single model to predict both the intent and slot labels simultaneously. The model uses the context of the input sentence to predict these labels. Joint models have been shown to achieve SOTA performance on several spoken language understanding datasets. Xu and Sarikaya (2013) propose a joint convolutional neural network (CNN) and RNN model for intent detection and slot filling on the ATIS (Upadhyay et al., 2018) dataset. They achieved SOTA results at the time. In the same context, Liu and Lane (2016) proposed an attention-based neural network for joint intent detection and slot filling. The model uses an attention mechanism to weigh the importance of different parts of the input sentence for predicting the intent label and slot labels. Chen et al. (2019) explore the use of the BERT model for joint intent detection and slot filling on ATIS (Upadhyay et al., 2018) and SNIPS (Coucke et al., 2018). They report SOTA results on both datasets.

3 SID4LR Shared Task

Task Formulation. Intent detection and slot-filling are critical NLP tasks where, given an utterance, a system is responsible for parsing the user's intent and extracting relevant information to act or reply appropriately. While many neural-based models have achieved SOTA performance for these tasks, their success often depends on large amounts of labeled data. However, many real-world datasets are limited to specific domains and are only available in English or a few other languages. As a result, it is important to reuse existing data in high-resource languages to develop models for low-resource languages, especially since tasks like intent classification and slot-filling require abundant labeled data.

Shared Task Problem Statement. This shared task of SID aims to address the challenges of performing SID for low-resource language varieties for the following languages: Swiss German, South Tyrolean, and Neapolitan. The training data provided consists of the Cross-lingual Slot and Intent Detection ($xSID_{0.4}$) corpus (van der Goot et al., 2021), a cross-lingual spoken language understanding dataset, covering 12 languages (Arabic, Chinese, Dutch, Danish, English, German, Indonesian, Italian, Japanese, Kazakh, Serbian, Turkish) from six language families with English training. The task allowed the use of pre-trained models and external data including data from the target language.

Evaluation Metric. The primary evaluation metric for slot filling is the span F_1 score, where both span and label must match exactly, and accuracy is used to evaluate intent detection where it is calculated through the ratio of the number of correct predictions of intent to the total number of sentences. More details regarding the shared task can be found in Aepli et al. (2023).

4 Data

Shared Task Data. The $xSID_{0.4}$ (van der Goot et al., 2021) corpus comprises cross-lingual SLU evaluation datasets covering 13 languages from six language families. The training dataset contains $43,605$ sentences, the development set contains 300 sentences, and the test set contains 500 sentences. The corpus contains sentences from Snips and Facebook, which were translated into all 13 target languages, resulting in a cross-lingual SLU evaluation dataset. All examples are annotated with

Language	# Train	# Valid	# Test
ar	$42,157$	300	500
da	$43,605$	300	500
de	$43,605$	300	500
en	$43,605$	300	500
id	$42,157$	300	500
it	$43,605$	300	500
ja	$29,073$	150	250
kk	$42,157$	300	500
nl	$43,605$	300	500
sr	$43,605$	300	500
tr	$43,605$	300	500
zh	$42,157$	300	500

Table 3: Number of samples in the train, validation, and test sets for each language in the multilingual dataset xSID$_{0.4}$, where the language codes are represented by two-letter ISO codes. The dataset includes 12 languages: Arabic (ar), Danish (da), German (de), English (en), Indonesian (id), Italian (it), Japanese (ja), Kazakh (kk), Dutch (nl), Serbian (sr), Turkish (tr), and Chinese (zh).

their intent and corresponding slots. Listing 1 provides examples of annotations with intent and slots. We converted the dataset into a JSON format that includes intents and slots. This JSON file was then converted to HuggingFace Dataset format for easy use with our transformer models. A sample of the resulting JSON format is shown in Listing 2.

```
# text: show all reminders
# intent: reminder/show_reminders
# slots: 5:8:reminder/reference,
         9:18:reminder/noun
1 show      reminder/show_reminders O
2 all       reminder/show_reminders B-reference
3 reminders reminder/show_reminders O
```

Listing 1: Example of the dataset format

```
{'text':'show all reminders',
 'slots': 'reference:all',
 'intent': 'reminder/show_reminders',
 '__index_level_0__': 0}
```

Listing 2: Example of the preprocessed dataset

External Data. As mentioned, Swiss German, South Tyrolean, and Neapolitan are low-resource languages with limited available labeled data. To address this challenge, we incorporate unlabeled data from different sources to augment our training data. We describe these external sources next.

SwissCrawl (Linder et al., 2020), a corpus of over $500,000$ Swiss German sentences gathered from web crawling between September and November 2019. The sentences are representative of how native speakers write in forums and social media and may contain slang and ascii emojis.

DiDi Corpus (Frey et al., 2016) is a multilingual language corpus of $600,000$ tokens from Facebook users in South Tyrol, Italy. It includes CMC texts, socio-demographic data, and linguistic annotations on thread, text, and token level. The corpus is mainly German and Italian, with English also present, and has been manually anonymized and annotated.

OSCAR Corpus (Caswell et al., 2021) is a large multilingual corpus created by scraping the web and includes texts in more than 200 languages. The OSCAR Corpus includes texts in Neapolitan, which is a Romance language spoken in the southern part of Italy, particularly in the region of Campania. The Neapolitan texts in the corpus consist of around 4.4 million tokens, making it one of the largest resources available for this language.

5 Pre-trained Language Models

In this study, we evaluate several popular multilingual Transformer-based language models, including mBERT, XLM-R, SBERT, LaBSE, LASER, and mT0. These models are capable of effectively capturing cross-lingual embeddings, enabling transfer learning across multiple languages. Below we provide a description of each model used in our experiments on the training dataset.

mBERT. is the multilingual version of BERT (Devlin et al., 2019), which is an encoder model with bidirectional representations from Transformers trained with a denoising objective. mBERT is trained on Wikipedia for 104 languages including German and Italian.

XLM-R. (Conneau et al., 2020) is a transformer-based multilingual masked language model pre-trained on more than 2TB of filtered CommonCrawl data in 100 languages, including languages including German and Italian. XLM-R uses a Transformer model (Vaswani et al., 2017) trained with a multilingual masked language model XLM (Conneau and Lample, 2019).

sBERT. Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), is a modification of the pretrained

BERT (Devlin et al., 2019) model that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity. As we work under a multilingual context, we use the multilingual versions from previously monolingual SBERT models (Reimers and Gurevych, 2020) which is trained for sentence embedding in 50+ languages from various language families.

LaBSE. Language-agnostic BERT Sentence Encoder (LaBSE) (Feng et al., 2020a) is a BERT-based model trained to generate sentence embeddings in 109 different languages. The model's pre-training approach involves a combination of masked language modeling and translation language modeling. The pre-training process combines masked language modeling with translation language modeling. LaBSE is useful for producing sentence embeddings in multiple languages and performing bi-text retrieval.

LASER. Language-Agnostic Sentence Representations (LASER) (Feng et al., 2020b) is a contextualized language model based on a BiLSTM encoder trained on parallel data from OPUS website (Tiedemann, 2012) using a translation objective. The LASER model can handle 200 different languages.

mT0. (Muennighoff et al., 2022) is a group of sequence-to-sequence models that come with different sizes from 300M to 13B parameters trained to investigate the cross-lingual generalization through multitask finetuning. mT0 can execute human instructions in many languages without any prior training. The models are fine-tuned from pre-existing mT5 (Xue et al., 2020) multilingual language models using a cross-lingual task mixture called xP3. These refined models are capable of cross-lingual generalization to unseen languages.

6 Experiments and Settings

Training on English Data. As a baseline setting, we train all the pre-trained models described in Section 5 on the English part of the multilingual dataset $\text{xSID}_{0.4}$ (van der Goot et al., 2021) and evaluate them on Swiss German, South Tyrolean, and Neapolitan under a zero-shot setting.

Training on German/Italian Data. Our second approach aims to train all the pre-trained models on the language family of low-resource languages (i.e., German for Swiss German and South Tyrolean, and Italian for Neapolitan, respectively) under the zero-shot setting. So, we extract the German and Italian SID data from $\text{xSID}_{0.4}$, and then fine-tune all our models on both datasets. Then, we evaluate the German models on GSW and ST tasks and the Italian models on the NAP task.

Training on Multilingual Dataset. Next, we explore a third training approach that involves the full multilingual $\text{xSID}_{0.4}$ dataset. To do so, we combine all the 12 available languages in the $\text{xSID}_{0.4}$ dataset and fine-tune our pre-trained models on this combined dataset. We then evaluate each target using a zero-shot setting. This approach allows us to train on larger and more diverse datasets. In total, we generate $502,936$ training sentences across all languages in the dataset.

Paraphrase and Machine Translation. To improve the performance of our pretrained models, we also explore the impact of data augmentation techniques such as paraphrasing and machine translation. Specifically, we aime to examine how these techniques can enhance the performance of our models on cross-lingual SLU tasks. To this end, we experiment with different data augmentation strategies, including paraphrasing and machine translation. Paraphrasing is performed using the quality-guided controlled paraphrase generation (QCPG) model (Bandel et al., 2022), resulting in a total of $130,815$ sentences in English. These sentences are then translated into German and Italian using the OPUS-MT model (Tiedemann and Thottingal, 2020), creating cross-lingual datasets for our experiment.

To further augment our training data for low-resource languages, we leverage Meta AI's No Language Left Behind (NLLB), which provides open-source models capable of high-quality translations between 200 languages (including low-resource languages (NLLB Team et al., 2022)). To create our new training data using the NLLB model, we first use FastText to detect the language codes of our target languages. Next, we utilize NLLB models to translate the English training data into the predicted language codes. The language codes identified for our target languages are *deu_latn* for Swiss German, *est_latn* for South Tyrolean, and *ita_Latn* for Neapolitan. We generate $43,605$ sentences for each of the three languages. It is worth noting that we ensure that the labels for each sentence remain the same throughout the paraphrasing and machine translation process to maintain the integrity of the data.

Training on External Data. Since the language models we employ do not have a strong representation of the low-resource languages used on the task, we leverage large corpora of each of the low-resource languages into the training process. By incorporating external datasets, the models are exposed to more comprehensive information about the semantics of each low-resource language, enabling them to better capture the nuances and complexities of the target languages.

Training MT0 As discussed in Section 5, the MT0 models share the same architecture as MT5/T5 models, i.e., they are encoder-decoder models. Therefore, we train them for intent classification and slot detection using the data preprocessing approach described in Section 4. We utilize the PEFT library provided by Huggingface (Sourab Mangrulkar, 2022) to train the mT0-small, mT0-Base, and mt0-Large models. Our approach involves using LORA (Hu et al., 2021), which allows us to achieve SOTA performance while consuming significantly less memory. For the mT0-xxl models, we utilize DeepSpeed (Rasley et al., 2020) with CPU offloading to train a model with 13B parameters on a 40GB A100 GPU.

Combining Models. In recent studies, joint learning techniques that combine multiple classification approaches have produced promising results (Bilat et al., 2020). These approaches involve concatenating the outputs of individual models and passing the resulting output through multiple neural network layers, allowing the resulting network to be trained jointly. In this part of our experiments, we investigate the effectiveness of this approach in zero-shot settings by combining multilingual models. Specifically, we combine LASER embeddings, from the LASER model, with other multilingual models including mBERT, sBERT, LaBSE and XLMR.

7 Results and Discussions

Evaluation on Validation Data. We present the accuracy scores of all our models across various settings. Table 4 presents the evaluation results for the intent classification task on the validation set. Our transformer-based models, with different experimental settings, outperform the baseline on all the target languages. For instance, **mT0-base** outperforms the baseline (mBERT) with an average of +16.49, +22.93, +17.90 for GSW, ST, and NAP, respectively. Notably, our best combination was the mT0-xxl model under the multilingual setting. It achieves the best results of 89.00, 94.00, and 87.00, improving the baseline with +29.30, +33.30, and +25.70 Accuracy point in the three target languages.

The results of the slot filling task on the validation set are shown in Table 5. Our transformer-based models perform better than the baseline across all target languages when tested under different experimental settings. Our best-performing model, mT0-large, achieves the most outstanding results using the Multilingual settings with F_1 scores of 60.30, 55.00, and 52.30 in the three target languages. These results represent a notable improvement over the baseline, with an increase of +30.88, +4.65, and +0.90 F_1 points in the three target languages.

Our results on the validation data suggest that larger models generally achieve better performance, implying that higher parameter counts result in better cross-lingual and zero-shot setting performance. Moreover, as the mT0 models are fine-tuned from pre-existing mT5 multilingual language models, they are capable of performing cross-lingual generalization on unseen languages. This capability may be a possible reason for the mT0 models outperforming other models in zero-shot settings.

Official Shared Task (Test) Results. Our findings regarding the performance of larger models are also observed in the test set. Table 6 presents the evaluation results for both slot filling and intent classification tasks across all three target languages. Our mT0 models strongly outperform the baseline models. Specifically, our mT0 models outperformed the baseline models in all target languages for the intent classification task, highlighting the effectiveness of larger models for intent classification. Moreover, our mT0 models also outperform the baseline models in two of the target languages for slot filling task, further indicating the superiority of larger models for sentence-level classification tasks. The improvement in scores for intent classification is more evident than for slot filling. The larger improvement in scores for intent classification may be correlated with the fact that for our data augmentation experiment on paraphrasing and machine translation, we were only able to augment data for intent classification, resulting in a larger improvement in performance for this task compared to slot filling.

It is worth noting that we use the validation set for model selection, which resulted in higher scores than those achieved on the test set. This is because

Setting	Lang.	mBert	LS	LL	LX	mT0-small	mT0-base	mT0-large	mT0-xxl
English	GSW	51.67	45.30	52.70	48.30	69.20	70.20	69.00	<u>80.00</u>
	ST	61.00	58.00	66.70	61.70	74.50	76.20	79.10	<u>89.00</u>
	NAP	61.00	55.30	56.00	67.0	71.30	72.00	75.00	<u>76.33</u>
German	GSW	59.00	74.00	68.00	80.70	69.30	73.33	80.33	<u>84.33</u>
	ST	59.70	55.70	59.00	51.00	83.33	88.33	84.66	<u>92.00</u>
Italian	NAP	65.30	63.30	63.70	55.70	77.66	84.66	83.33	<u>86.00</u>
Multilingual	GSW	59.70	62.70	59.70	53.30	75.00	76.33	84.00	<u>89.00</u>
	ST	60.70	54.70	58.00	56.30	88.33	85.66	90.66	<u>94.00</u>
	NAP	61.30	55.70	59.00	60.30	82.66	84.66	86.00	<u>87.00</u>
Paraphrase+MT	GSW	45.30	37.30	58.00	64.00	79.00	83.00	84.33	**91.00**
	ST	61.70	61.30	60.70	60.70	90.66	93.00	90.00	**95.66**
	NAP	63.70	60.00	58.70	60.00	85.66	**89.00**	87.33	88.33

Table 4: Accuracy results for intent classification on the validation set. **Baseline:** mBERT (Devlin et al., 2019). **LS:** LASER (Feng et al., 2020b)+sBERT (Reimers and Gurevych, 2019). **LL:** LASER+LaBSE (Feng et al., 2020a). **LX:** LASER+XLM-R (Conneau and Lample, 2019). **<u>Underline</u>:** Best-performing models for each setting. **Bold:** Best F_1 score across all the experiments and settings.

Setting	Lang.	Baseline	mt0-small	mt0-base	mt0-large
English	GSW	26.23	25.42	34.00	**40.32**
	ST	44.61	32.40	44.00	**54.30**
	NAP	48.01	42.20	47.90	**49.00**
Multi-langl	GSW	29.42	28.90	42.30	**60.30**
	ST	50.35	43.40	53.40	**55.00**
	NAP	51.40	49.00	50.30	**52.30**

Table 5: Slot-f1 results for Slot Filling on the validation set. Bold entries are the best-performing models for each experiment and setting.

Task	Lang.	Baseline	mT0-large
Slots	ST	44.61	**46.41**
	GSW	26.23	**27.39**
	NAP	**48.01**	38.82
Intents	ST	61.00	**89.40**
	GSW	51.67	**81.60**
	NAP	61.00	**85.40**

Table 6: Results on the test set for both SID tasks. **Bold** entries indicate the model's performance compared to the baseline model.

the validation data is similar to the data used during training, while the test data is entirely new and unseen. As a result, the test scores may be lower due to differences in the distribution of data between the training and test sets. Nevertheless, our mT0 models consistently outperform the baseline models on the test set, providing further evidence for the effectiveness of larger models in SID tasks.

8 Conclusion

We described our contribution to the SID4LR (Aepli et al., 2023) shared tasks. Our models target both the slot and intent sub-task in three proposed low-resource languages, namely, Swiss German, South Tyrolean, and Neapolitan. We test the utility of existing pretrained language models such as mT0 (Muennighoff et al., 2022) on the intent detection and slot filling tasks. We show that such models can lead to improving the results of the baseline with an average of +27 F_1 points. In the future, we intend to use mT0 to jointly model the intent detection and slot filling tasks for improving overall performance.

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 *Proceedings of the Tenth Workshop on NLP for Similar Languages, Varieties and Dialects*, Dubrovnik, Croatia. Association for Computational Linguistics.

Elron Bandel, Ranit Aharonov, Michal Shmueli-Scheuer, Ilya Shnayderman, Noam Slonim, and Liat Ein-Dor. 2022. Quality controlled paraphrase generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 596–609, Dublin, Ireland. Association for Computational Linguistics.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. SLURP: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7252–7262, Online. Association for Computational Linguistics.

Loïs Bilat, Georgiana Dinu, and Mark Cieliebak. 2020. Cross-lingual toxicity detection. pages 50–55.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Javier Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2021. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *arXiv e-prints*, page arXiv:2103.12028.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. BERT for joint intent classification and slot filling. *CoRR*, abs/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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

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.

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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020a. Language-agnostic bert sentence embedding.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020b. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

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. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Jennifer Frey, Aivars Glaznieks, and Egon Stemle. 2016. The didi corpus of south tyrolean cmc data: A multilingual corpus of facebook texts.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. 2018. Semantic parsing for task oriented dialog using hierarchical representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2787–2792, Brussels, Belgium. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

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.

Xiujun Li, Yu Wang, Siqi Sun, Sarah Panda, Jingjing Liu, and Jianfeng Gao. 2018. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems.

Lucy Linder, Michael Jungo, Jean Hennebert, Claudiu Cristian Musat, and Andreas Fischer. 2020. Automatic creation of text corpora for low-resource languages from the internet: The case of swiss german. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2706–2711, Marseille, France. European Language Resources Association.

Bing Liu and Ian R. Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *CoRR*, abs/1609.01454.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking natural language understanding services for building conversational agents.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech Model Pre-Training for End-to-End Spoken Language Understanding. In *Proc. Interspeech 2019*, pages 814–818.

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, and Geoffrey Zweig. 2015. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):530–539.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

P. J. Price. 1990. Evaluation of spoken language systems: the ATIS domain. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

Suman Ravuri and Andreas Stolcke. 2015. Recurrent neural network and lstm models for lexical utterance classification. pages 135–139.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

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.

Lysandre Debut Younes Belkada Sayak Paul Sourab Mangrulkar, Sylvain Gugger. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021. 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Puyang Xu and Ruhi Sarikaya. 2013. Convolutional neural network based triangular crf for joint intent detection and slot filling. pages 78–83.

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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Kaisheng Yao, Baolin Peng, Geoffrey Zweig, Dong Yu, Xiaolong Li, and Feng Gao. 2014. Recurrent conditional random field for language understanding. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4077–4081.

Jianguo Zhang, Kazuma Hashimoto, Yao Wan, Ye Liu, Caiming Xiong, and Philip S. Yu. 2021. Are pre-trained transformers robust in intent classification? A missing ingredient in evaluation of out-of-scope intent detection. *CoRR*, abs/2106.04564.