Data Augmentation and Learned Layer Aggregation for Improved Multilingual Language Understanding in Dialogue

Evgeniia Razumovskaia Ivan Vulić Anna Korhonen
Language Technology Lab, University of Cambridge
{er563, iv250, alk23}@cam.ac.uk

Abstract

Scaling dialogue systems to a multitude of domains, tasks *and* languages relies on costly and time-consuming data annotation for different domain-task-language configurations. The annotation efforts might be substantially reduced by the methods that generalise well in zero- and few-shot scenarios, and also effectively leverage external unannotated data sources (e.g., Web-scale corpora). We propose two methods to this aim, offering improved dialogue natural language understanding (NLU) across multiple languages: **1)** *Multi-SentAugment*, and **2)** *LayerAgg*. Multi-SentAugment is a self-training method which augments available (typically few-shot) training data with similar (automatically labelled) in-domain sentences from large monolingual Web-scale corpora. LayerAgg learns to select and combine useful semantic information scattered across different layers of a Transformer model (e.g., mBERT); it is especially suited for zero-shot scenarios as semantically richer representations should strengthen the model's cross-lingual capabilities. Applying the two methods with state-of-the-art NLU models obtains consistent improvements across two standard multilingual NLU datasets covering 16 diverse languages. The gains are observed in zero-shot, few-shot, and even in full-data scenarios. The results also suggest that the two methods achieve a synergistic effect: the best overall performance in few-shot setups is attained when the methods are used together.

1 Introduction

The aim of Natural Language Understanding (NLU) in task-oriented dialogue systems is to identify the user's need from their utterance (Xu et al., 2020). This comprises the following crucial information: **1)** *intents*, what the user intends to do, and **2)** (typically predefined) *slots*, associated arguments of the intent (Tur et al., 2010; Tur and De Mori, 2011) which need to be filled with specific *values*. Intent detection is often framed as a standard sentence classification task, where every sentence maps to one or more intent classes; slot labelling is typically cast as a sequence labelling task, where each word is labelled with a BIO-style slot tag (Bunk et al., 2020), see Figure 1.

Figure 1: Illustration of two user utterances in the ATIS flight domain with associated intents and slot tags.

The supervised models for NLU in English are plentiful and achieve extremely high accuracy (Louvan and Magnini, 2020a; Qin et al., 2021). At the same time, porting an NLU system to any new domain *and* language requires collecting a large in-domain dataset, and training a model for the target language (Xu et al., 2020). Such in-domain annotations in multiple languages are extremely expensive and time-consuming (Rastogi et al., 2020), also reflected in the fact that large enough dialogue NLU datasets for other languages are still few and far between (Razumovskaia et al., 2021). This in turn creates the demand for strong multilingual and cross-lingual methods which generalise well and learn effectively in zero-shot and few-shot scenarios. In this work, we propose two methods to this end: **1)** *Multi-SentAugment*, a weakly supervised data augmentation method which improves the capability of current state-of-the-art (SotA) dialogue NLU in few-shot scenarios via self-training; **2)** *LayerAgg* learns to effectively leverage and combine the knowledge stored across different layers of a pre-trained multilingual Transformer (e.g., mBERT).

The main goal of *Multi-SentAugment* is to reduce the required amount of labelled data and manual annotation labour by harvesting the large pool

Findings of the Association for Computational Linguistics: ACL 2022, pages 2017 - 2033

of unannotated data, and carefully selecting relevant in-domain examples which can then be automatically labelled (Du et al., 2021). In a nutshell, domain-relevant unannotated sentences are first retrieved from a large multilingual sentence bank. The *synthetic* labels for the data are then generated by a teacher model, previously trained with available annotated data. A final student model is then trained on the combination of synthetically labeled and annotated data. To the best of our knowledge, our work is the first to mine large unannotated monolingual resources in multiple languages to augment data for multilingual dialogue NLU.

The goal of *LayerAgg* is to leverage useful lexical and other semantic information scattered across layers (Tenney et al., 2019; Vulić et al., 2020) of a pretrained multilingual Transformer. Moving away from the standard fine-tuning practice of using only the representations from the top layer, we hypothesise that the model's cross-lingual capabilities can be increased by forcing it (i) to propagate semantic information from lower layers, as well as (ii) to aggregate/combine semantic information from all its layers. In a nutshell, we propose to use a *multilingual encoder with cross-layer Transformer*, which selects and combines the knowledge from all layers of a pretrained model during fine-tuning.

Our experiments show that Multi-SentAugment gives consistent improvements in few-shot and full-data scenarios on the two available multilingual dialogue NLU datasets: MultiATIS++ (Xu et al., 2020) and xSID (van der Goot et al., 2021). The results further indicate that LayerAgg improves zero-shot performance on the same datasets. Finally, since the two methods can be independently applied to SotA NLU models, we demonstrate that they yield a synergistic effect: the highest scores on average are achieved with their combination.

Contributions. 1) Multi-SentAugment is a simple yet effective data augmentation approach which leverages unannotated data from large Web-scale corpora to boost multilingual dialogue NLU. **2)** LayerAgg is a novel cross-layer attention method which learns to effectively combine useful semantic information from multiple layers of a multilingual Transformer. **3)** The two methods applied with SotA NLU models obtain consistent gains across two standard multilingual NLU datasets in zero-shot, and 8 languages in few-shot, and full-data setups, boosting the capability of cross-lingual dialogue in resource-lean scenarios.

2 Related Work and Background

Multilingual NLU for Dialogue Systems is usually divided into two tasks: intent detection and slot labelling (Tur et al., 2010; Xu et al., 2020). In "pre-Transformer" times, the methods for training multilingual NLU systems were based on static multilingual word vectors (Mrkšić et al., 2017; Upadhyay et al., 2018; Schuster et al., 2019), lexicon alignment (Liu et al., 2019b,a), and model or annotation projection via parallel data (Kulshreshtha et al., 2020; López de Lacalle et al., 2020).

Transfer learning with large pretrained multilingual Transformer-based language models (LMs) such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020a) has demonstrated currently unmatched performance in many NLU tasks (Liang et al., 2020; Hu et al., 2020; Ponti et al., 2020; Ruder et al., 2021), including intent classification and slot labelling (Zhang et al., 2019; Liu et al., 2020). Fine-tuning a large multilingual LM has become a standard for multilingual NLU (Zhang et al., 2019; Xu et al., 2019; Kulshreshtha et al., 2020). However, the excessively high data annotation costs for multiple domains and languages still hinder progress in multilingual dialogue (Razumovskaia et al., 2021). In this paper, unlike prior work, we propose to use external unannotated data to mine and automatically label in-domain in-language examples which aid learning in low-data regimes across multiple languages.

Data Augmentation in Multilingual NLU, as well as data augmentation methods in NLP in general, aim to produce additional training data automatically, without the need to manually label it. In monolingual English-only settings, English NLU data has been augmented by generating additional data with a large monolingual language model (Peng et al., 2020) such as BERT (Devlin et al., 2019) or GPT-2 (Radford et al., 2019), or from atomic templates (Zhao et al., 2019). In multilingual settings, data augmentation methods for NLU include simple text span substitution and syntactic structure manipulation (Louvan and Magnini, 2020c,b). Recently, code switching (Krishnan et al., 2021) and generating translations through a pivot language (Kaliamoorthi et al., 2021) have also been proposed as data augmentation methods.

The previous work relies on (i) additional components such as syntactic parsers or POS taggers, or (ii) parallel and code-switched data. However, they might be unavailable or of low-quality for

many (low-resource) languages. In contrast, Multi-SentAugment relies on the cheapest and largest resource available: monolingual Web-crawled data; it disposes of any dependency parsers and taggers, which makes it more widely applicable. Mining knowledge from Web-scale data was shown effective in various (non-dialogue) text classification tasks (Du et al., 2021) and in MT (Wu et al., 2019).[1]

Layer Aggregation in Pretrained LMs. A standard practice is to use the output of the final/top layer of a pretrained LM as input into task-specific classifiers (Devlin et al., 2019; Sun et al., 2019). At the same time, prior work shows that most of (decontextualised) lexical information (Ethayarajh, 2019; Vulić et al., 2020) and word-order information (Lin et al., 2019) is localised in lower layers of BERT. Middle layers usually encode syntactic information (Hewitt and Manning, 2019; Jawahar et al., 2019) while (contextual) semantic information is spread across all the layers of a pretrained LM (Tenney et al., 2019), with higher layers capturing increasingly abstract language phenomena (Lin et al., 2019; Rogers et al., 2020; Tenney et al., 2019). Kondratyuk and Straka (2019) showed that using a weighted combination of all layers works well in cross-lingual settings for a syntactic task of dependency parsing. In addition, they proposed to use layer dropout to redistribute how the information is localised in a fine-tuned BERT model.

In order to 'unlock' additional semantic knowledge from other layers, we propose an additional Transformer encoder with cross-layer attention as a layer aggregation mechanism. We hypothesise that relying only on the representations from the top layer dilutes mBERT's lexical and semantic information. Moreover, we expect lexically and semantically richer representations to be especially useful for zero-shot settings: aggregated (contextualised) semantic information from lower layers could help correctly identify the intent of the sentence, while lexical information could help identify the slot tag for different languages.[2]

3 Methodology

We assume a standard state-of-the-art approach to dialogue NLU in multiple languages (Xu et al., 2020), based on fine-tuning pretrained multilingual LMs on the tasks of intent detection and slot labelling. Following Xu et al. (2020), we fine-tune the pretrained LM in a standard supervised fashion, with task-specific linear layers stacked on top.

Separate NLU Models. The multilingual encoder for each NLU task is fine-tuned separately, and there is no knowledge exchange (but also no noise or destructive inference) between the two tasks. We adopt a standard task-specific fine-tuning setup (Xu et al., 2020; Siddhant et al., 2020).

Joint NLU Model. Another line of recent work pursued joint modelling of the two tasks, motivated by the intuitive correlation between them.[3] In this work, we follow a standard joint modelling procedure (Xu et al., 2020; Hardalov et al., 2020; Krishnan et al., 2021), where the model consists of a shared multilingual encoder followed by task-specific linear layers for intent classification and slot labelling. The loss is then simply a sum of two task-dedicated losses. In our experiments, we use mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020a) as the encoder.

Multi-SentAugment (§3.1) and LayerAgg (§3.2) are then applied to the joint NLU model, while we also provide detailed comparisons to the separate NLU models as baselines in zero-shot setups.

3.1 Multi-SentAugment

Large Web-crawled datasets have been proven useful for extracting additional data for classification tasks in English (Du et al., 2021). We adapt the approach of Du et al. (2021) to multilingual dialogue NLU, that is, we propose to use large Web-crawled corpora to obtain additional in-domain data for dialogue NLU tasks in multiple languages.

For each language l we are given: 1) some annotated training data D_l which consists of $|D_l|$ sentences $x_1, ..., x_{|D_l|}$, each labelled with intent class and slot labels (see Figure 1); 2) a large Web-crawled corpus U_l consisting of $|U_l|$ sentences $s_1, ..., s_{|U_l|}$; 3) off-the-shelf multilingual sentence encoder $\mathbb{F}$ fine-tuned towards semantic sentence similarity, that is, to produce semantic embeddings of input sentences (Reimers and Gurevych, 2020). The data augmentation process then consists of 1) unsupervised data retrieval and 2) self-training.

[1] Unlike Du et al. (2021), we do not tune pretrained language models to sentence similarity, but use off-the-shelf pretrained multilingual sentence encoders (Artetxe and Schwenk, 2019; Feng et al., 2020; Litschko et al., 2021).

[2] For instance, *10.07.2021* will be typically identified as *date* in many languages.

[3] Information about the slots in an utterance could be informative of its intent, and vice versa. For instance, an utterance containing `temperature unit` slot is more likely to belong to intent `find_weather` than to intent `set_alarm`.

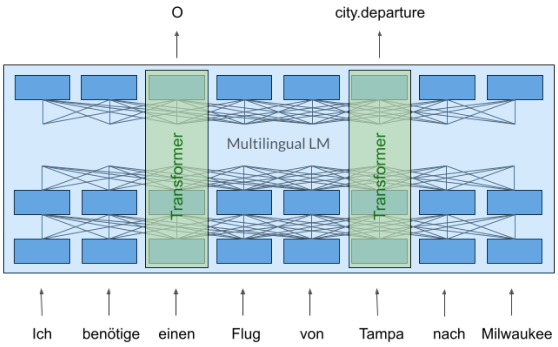


Figure 2: Illustration of the LayerAgg method.

The aim of *unsupervised data retrieval* is to construct an in-domain unannotated set of sentences by filtering the sentences from U_l. The process is formulated by the following equations:

$$\mathbf{X} = \mathbb{F}(x_1, \ldots, x_{|D_l|}); \mathbf{U} = \mathbb{F}(s_1, ..., s_{|U_l|});$$
$$\sigma = \frac{\mathbf{U}\mathbf{X}^\top}{\|\mathbf{U}\|\|\mathbf{X}\|} > \theta;$$

θ is a similarity threshold for sentence filtering: a sentence s_i will be added into the in-domain dataset if there is an annotated sentence $x_j \in D_l$ such that $\sigma_{i,j} > \theta$. As a result of data retrieval, we obtain a set of in-domain unannotated sentences which are similar to annotated training data D_l.

At *self-training*, we first fine-tune a joint NLU model on annotated D_l data. We then use this model to annotate the retrieved in-domain sentences. As our final NLU model, we fine-tune a new joint NLU model on the full dataset, combining the D_l set and filtered and annotated sentences.

3.2 LayerAgg

To ensure the propagation and use of lexical and semantic information from lower layers, we propose a simple layer aggregation technique based on cross-layer attention (Vaswani et al., 2017), illustrated in Figure 2. In short, let $\mathbf{w}_{ij}$ be a representation of a word (or WordPiece; Devlin et al. (2019)) at position i at layer j, $j = 1, \ldots, N_l$, where N_l is the number of layers in the pretrained LM (e.g., $N_l = 12$ for mBERT). Layer-aggregated representation $\mathbf{w}_i$ of the input w_i is computed as follows:

$$\mathbf{w}_i = \mathbb{T}(\mathbf{w}_{i,1:N_l}), \quad (1)$$

where $\mathbf{w}_{i,1:N_l}$ is a sequence comprising all (ordered) $\mathbf{w}_{ij}$ per-layer representations, and $\mathbb{T}$ is a cross-layer Transformer encoder. In essence, $\mathbb{T}$ effectively always operates over a sequence of length N_l: it outputs the representations from all layers, but which have now been self-attended. We then feed the last item (i.e., N_l-th item) of the sequence representation output by the Transformer $\mathbb{T}$ into the task-specific classifiers. Relying on the N_l-th output representation, the model is forced to incorporate the information from all layers into the final representation of the input token w_i. The parameters of $\mathbb{T}$ are also updated during fine-tuning.

Dataset	Languages	Utterances	Intents	Slots
MultiATIS++	de, en, es, zh, ja, fr, pt	5871	18	84
	tr	1353	17	71
	hi	2493	17	75
xSID	en	43605	13	16
	ar, da, de, de-st (st), id, it, kk, nl, sr, tr, zh	800		
	ja	400		

Table 1: Dataset statistics for MultiATIS++ and xSID. Language codes are available in the Appendix.

4 Experimental Setup

Evaluation Datasets comprise two standard multilingual dialogue NLU datasets: MultiATIS++ (Xu et al., 2020) and xSID (van der Goot et al., 2021), created by translating monolingual labelled English data into target languages. MultiATIS++ is a single domain (airline) dataset while xSID covers 7 domains including *alarm, weather, music, events* and *reminder*. xSID is an evaluation only dataset, i.e., it contains training data only for English. The statistics of the datasets are presented in Table 1. The datasets consist of sentences each labelled with an intent class and BIO slot tags/labels, see Figure 1.

Large (Multilingual) Sentence Banks. We use the CC-100 dataset (Conneau et al., 2020a; Wenzek et al., 2020), which comprises monolingual CommonCrawl data in 116 languages. For computational tractability with resources at our disposal, we rely on the smaller CC-100-100M dataset, a random sample from the full CC-100[4] spanning 100M sentences in each language. CC-100 covers multiple domains, language styles and variations.

Multi-SentAugment: Setup. Unless noted otherwise, we use the LASER multilingual sentence encoder (Artetxe and Schwenk, 2019), pretrained on

[4] http://data.statmt.org/cc-100/

93 languages with a sentence similarity objective on parallel data. The similarity threshold θ is set to 0.8. Besides the basic setup, (i) we also analyse the impact of the sentence encoder by running experiments with another SotA multilingual encoder: LaBSE (Feng et al., 2020; Litschko et al., 2021); (ii) we apply an additional filtering step based on the intent confidence of the teacher model, retaining only high-confidence examples.[5]

LayerAgg. The aggregator Transformer $\mathbb{T}$ contains a single 512-dimensional layer with 4 attention heads. Here, we remind the reader that the N_l-th item of $\mathbb{T}$'s output sequence is fed to the task-specific layers; see again §3.2). LayerAgg adds up to 2 million additional parameters, which is $\approx 1\%$ of the total number of trainable parameters in the baseline model. In addition, we present an extensive comparison with a standard layer aggregation method of Kondratyuk (2019), which is based on cross-layer attention.

Fine-Tuning Setup. 1) In the **zero-shot** setup, we train the model on the English training data and evaluate on other (target) languages. 2) In the **few-shot** setup, unless stated otherwise, we add 10 target-language examples (i.e., shots) per intent to the English training data. 3) In the **full-data** setup, we use the entire training set of the target language (without any English data). For unsupervised sentence retrieval in few-shot and full-data setups, we only use the examples in the target language as our query set D_l (see §3.1). In all experiments, we evaluate on the validation set after each epoch, and train for 20 epochs with a patience of 5 epochs, with Adam (Kingma and Ba, 2015) as the optimiser, batches of 32; the learning rate is $5e-5$, and the warm-up rate is 0.1. We experiment with mBERT Base and XLM-R Base as multilingual encoders. The hyperparameters were set to the values corresponding to those in Xu et al. (2020).

5 Results and Discussion

Joint vs Separate NLU. We first establish the performance of joint versus separate baseline NLU models. The main results, provided in Tables 2 and 3, indicate that joint NLU training performs better on intent classification while separate task-specific NLU models are more beneficial on slot labelling. Our results corroborate the findings from prior work (Schuster et al., 2019; He et al., 2020; Weld et al., 2021). We suspect that joint training works better for intent classification as sentence-level representations are enriched with lexical information through the additional slot-labelling loss. At the same time, separate training attains stronger performance in slot labelling as it retains more task-specific representations for each token.

Impact of LayerAgg. The motivation behind LayerAgg is to combine the strengths of both joint and separate training, that is, having sentence-level representations enriched with lexical information while keeping token representations specified. The benefits of LayerAgg in both tasks in zero-shot setups are indicated by the results in Tables 2-3. We observe large improvements with LayerAgg, both on average and for the large number of individual target languages. It is worth noting that LayerAgg provides gains also with both underlying multilingual encoders. Besides that, adding LayerAgg also yields more stable performance of the joint model in general (e.g., compare the scores on Japanese and Turkish slot labelling without and with LayerAgg). The gains with LayerAgg also persist in few-shot and full-data setups, as shown in Figure 3.

+LayerAgg versus +Attn. Table 2 also presents a comparison of two layer aggregation techniques: cross-layer attention from Kondratyuk and Straka (2019) (**+Attn**), now adapted to dialogue NLU tasks, and LayerAgg. While both methods produce gains over the **Joint** baseline in several target languages, LayerAgg yields much more substantial gains, and is more robust across different model configurations and tasks. While the Attn aggregation simply provides a weighted sum of information encoded across Transformer layers based on its importance to the final prediction, LayerAgg has the capability to analyse and aggregate the information as it evolves between layers (Voita et al., 2019).

Impact of Multi-SentAugment. The results in Figure 3 suggest that Multi-SentAugment is indeed useful as data augmentation for the two NLU tasks, both in few-shot and full-data scenarios, and for different target languages.[6] Achieving slight gains in full-data scenarios implies that mining additional monolingual data is beneficial even when a large in-domain dataset in the target language is avail-

[5] In practice, when we label extracted sentences with the teacher model, we only retain the sentences where the teacher model is confident in its prediction, that is, it assigns the intent class probability $p \geq 0.95$.

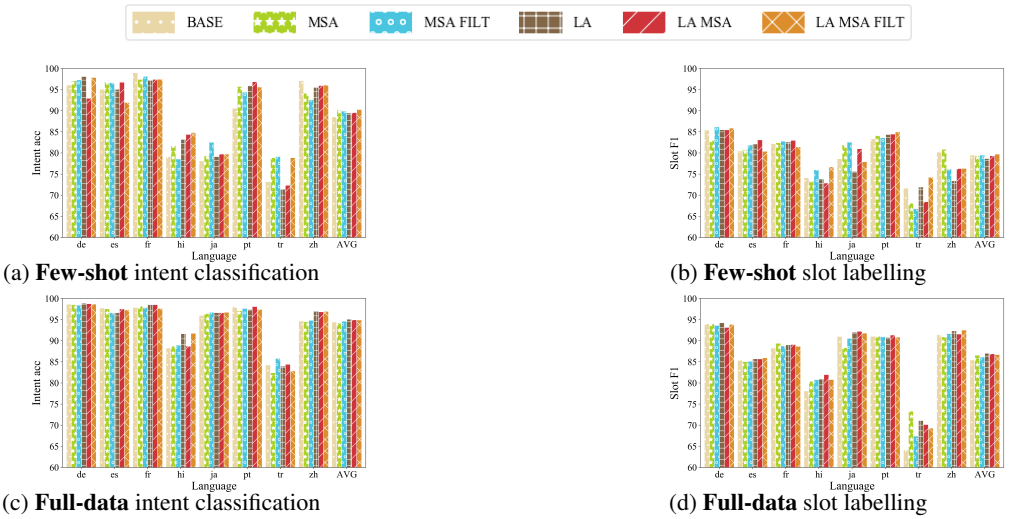
[6] We suspect that a slight performance drop in few-shot setups for `zh` and `ja` mostly stems from some discrepancy in tokenization between MultiATIS++ and CC-100.

Target language	de	en	es	fr	hi	ja	pt	tr	zh	AVG
Intent classification (Accuracy × 100)										
Separate mBERT	89.25	*98.66*	90.71	91.71	74.23	**77.27**	91.83	64.54	82.19	82.72
Joint mBERT	86.45	*98.54*	87.79	93.39	75.71	76.71	91.83	**70.78**	**84.55**	83.40
+Attn	85.67	*98.66*	88.91	87.57	**76.63**	**80.52**	91.04	69.65	84.21	83.02
+LayerAgg	**90.03**	*98.54*	**93.28**	**94.51**	74.92	77.27	**92.95**	70.21	81.52	**84.34**
Joint XLM-R	91.42	*98.45*	91.20	91.42	**80.99**	80.96	92.47	**71.94**	84.41	85.60
+Attn	91.12	*98.88*	90.41	91.12	78.01	82.16	94.79	70.56	83.32	85.19
+LayerAgg	**94.81**	*98.73*	**91.97**	**93.58**	78.28	**84.25**	**92.68**	68.41	**86.15**	**86.27**
Slot labelling (Slot F1 × 100)										
Separate mBERT	70.41	*95.20*	73.31	66.66	39.13	56.54	63.00	**49.31**	**56.65**	59.38
Joint mBERT	**70.52**	*95.54*	70.20	67.20	41.00	48.20	63.20	41.17	56.48	57.25
+Attn	70.14	*95.44*	70.48	**68.30**	**44.46**	52.89	**64.64**	48.20	56.46	59.46
+LayerAgg	69.15	*95.26*	**73.58**	**68.26**	43.59	**58.05**	**64.55**	48.08	55.62	**60.11**
Joint XLM-R	**81.57**	*95.58*	81.05	73.24	33.71	48.22	75.65	**38.92**	65.27	62.20
+Attn	79.88	*95.58*	80.40	70.50	33.20	46.45	75.33	38.60	65.62	61.25
+LayerAgg	80.93	*95.91*	**81.11**	**74.02**	**34.06**	**57.88**	**77.06**	**38.94**	**72.62**	**64.58**

Table 2: Zero-shot results on MultiATIS++ (English is the source language in all experiments). The average is computed across target languages (excluding English). Highest scores in each task for every encoder per column in **bold**. The results are averaged across 5 random seeds. **+Attn** refers to using standard cross-layer attention as layer aggregation, as done in prior work (Kondratyuk and Straka, 2019).

Target language	ar	da	de	st	en	id	it	ja	kk	nl	sr	tr	zh	AVG
Intent classification (Accuracy × 100)														
Joint mBERT	46.13	**74.07**	62.67	47.07	*98.80*	68.00	58.47	35.47	40.07	**65.87**	58.13	47.60	**72.61**	56.35
+LayerAgg	**51.13**	72.93	**63.00**	**49.47**	*98.67*	**69.00**	**62.20**	**39.33**	**47.53**	**65.73**	**61.73**	**50.80**	69.64	**58.54**
Joint XLM-R	51.07	86.40	70.73	48.20	*98.73*	81.87	69.13	39.60	45.53	79.20	70.07	72.00	77.60	65.95
+LayerAgg	**57.40**	**86.60**	**73.00**	**53.33**	*98.80*	**83.27**	**73.07**	**46.67**	**48.80**	**80.27**	**72.33**	**75.93**	**85.60**	**69.69**
Slot labelling (Slot F1 × 100)														
Joint mBERT	19.98	34.66	35.86	17.39	*95.37*	**29.45**	34.63	23.28	33.58	**38.37**	**25.74**	32.90	**63.80**	32.47
+LayerAgg	**21.00**	**36.21**	**37.97**	**18.51**	*94.27*	28.74	**35.50**	**30.19**	**35.58**	**38.91**	**25.79**	**35.32**	62.00	**33.77**
Joint XLM-R	32.40	**68.81**	**53.72**	20.68	*94.97*	64.31	**56.93**	**25.45**	**28.97**	**71.57**	48.96	46.78	56.42	47.91
+LayerAgg	**35.36**	68.50	52.16	**21.24**	*95.67*	**66.21**	**56.78**	23.68	**28.60**	68.10	**50.57**	**47.91**	**56.96**	**48.01**

Table 3: Zero-shot results on xSID. The average is computed across target languages (excluding English). Highest scores in each task for every encoder per column in **bold**. The results are averaged across 5 random seeds.



(a) **Few-shot** intent classification (b) **Few-shot** slot labelling

(c) **Full-data** intent classification (d) **Full-data** slot labelling

Figure 3: Few-shot and full-data results on MultiATIS++. BASE = joint training baseline; MSA = +Multi-SentAugment; MSA FILT = +Multi-SentAugment filtered by teacher model confidence; LA = +LayerAgg; LA MSA = +LayerAgg +Multi-SentAugment; LA MSA FILT = +LayerAgg +Multi-SentAugment filtered by teacher model confidence. Results are presented for mBERT, with same trends observed when using XLM-R. The full results for few-shot and full data scenarios are available in the Appendix C.

able. Notably, we observe larger gains for Turkish and Hindi in Figure 3d: it is expected due to the fact that MutiATIS++ contains a smaller number of sentences for `tr` and `hi` than for the other target languages. Finally, the impact of filtering by teacher confidence (see §3.1) is inconsistent for intent classification (i.e., it seems to be target language-dependent) while it improves the results for slot labelling on average. Encouraged by these insights, we will investigate more sophisticated in-domain sentence mining methods in future work.

Combining Multi-SentAugment and LayerAgg results in a synergistic effect, based on the additional slight gains observed in Figure 3 (the full results are available in the Appendix C, including the MultiSentAugment results in 5-shot and 20-shot setups in the Appendix D). This is expected as the two methods offer distinct enhancements of the base joint NLU model: (i) Multi-SentAugment includes more diverse sentences and lexical information into the training data (i.e., enhancement at the *input level*), while (ii) LayerAgg aims to select and combine semantic information spread across mBERT's layers (i.e., *feature-level* enhancement).

Zero-Shot vs Few-Shot. As discussed before, using Multi-SentAugment and LayerAgg seems to benefit the base NLU model both in low-data and full-data setups; we observe gains also in 5-shot and 20-shot setups (see Appendix D). Similar to other NLP tasks (e.g., named entity recognition, parsing, QA) (Lauscher et al., 2020), few-shot setups (e.g., even having only 5 examples per intent or ≈80 annotated sentences in total) yield huge benefits over zero-shot setups (see Table 4; compare the results in Table 2 and Figure 3). Our results provide another empirical proof calling for more modelling effort in more realistic few-shot cross-lingual transfer setups (Lauscher et al., 2020; Zhao et al., 2021) in future work. We also observe that the results in 10-shot setups when both Multi-SentAugment and LayerAgg are used are mostly on par with the results in 20-shot setups with the base NLU model. In general, this finding validates that the proposed methods can indeed reduce the manual annotation effort.

6 Analysis and Further Discussion

Target Language Analysis. While both Multi-SentAugment and LayerAgg are language-agnostic techniques *per se*, the actual transfer results also depend on the linguistic properties of the source and target languages. We thus aim to answer the following question: *Which languages benefit most from Multi-SentAugment and LayerAgg?* To this end, we study the correlations between zero-shot and few-shot transfer performance (i.e., gains over the joint baseline when using the two methods) and source-to-target language distance, which is based on the language vectors obtained from the URIEL typological database (Littell et al., 2017). Following Lauscher et al. (2020), we consider the following linguistic features: syntax (`SYN`), encoding syntactic properties; language family memberships (`FAM`) and geographic locations (`GEO`).

The results are shown in Table 5. `SYN` similarity has the highest correlation with zero-shot performance gains in both NLU tasks. We suspect that this might stem from LayerAgg's prop-

Shots (# of sentences)	Intent classification	Slot labelling
0 (0)	83.41	57.25
5 (81)	84.63	75.08
10 (153)	88.53	79.51
20 (270)	89.37	81.24
Full (4488)	94.43	85.42

Table 4: Impact of the amount of annotated examples in the target language. The results are averages across 8 target languages on MultiATIS++ (Xu et al., 2020) with the baseline Joint NLU model (with mBERT as the multilingual encoder).

Data setup	Task	Method	`SYN`	`FAM`	`GEO`
Zero-shot	Intent classification	LayerAgg	-0.9356	-0.5252	-0.6849
	Slot labelling	LayerAgg	0.6787	0.5392	-0.0509
Few-shot	Intent classification	LayerAgg	-0.1970	-0.2830	-0.1556
		Multi-SentAugment	0.2433	0.0497	-0.5229
		LayerAgg + Multi-SentAugment	0.5274	0.0192	-0.1298
	Slot labelling	LayerAgg	-0.4227	-0.3112	-0.9544
		Multi-SentAugment	-0.0032	0.4203	0.3934
		LayerAgg + Multi-SentAugment	0.1525	-0.1367	-0.6525

Table 5: Correlation between performance gains provided by each method (LayerAgg, Multi-SentAugment, and their combination) on MultiATIS++ and language distance scores between English as the source language and target languages, based on different typological features from URIEL (`SYN`, `FAM`, `GEO`).

	de	en	es	fr	hi	pt	tr	AVG
Joint	86.96	86.03	75.12	92.31	90.0	86.64	53.69	81.54
+LayerAgg	**97.83**	**97.53**	**83.19**	**95.33**	**91.16**	**89.48**	**58.49**	**87.57**

Table 6: F_1 scores in a lexical probe of detecting the 1,000 most frequent words on MultiATIS++.

erty to selectively aggregate information from multiple layers, which is easier to learn if the input sequences have similar syntactic structures. In simple words, LayerAgg might benefit more if similar information is found at similar places in the input sentences. FAM and GEO similarities are more correlated with gains in few-shot settings. This might be due to the fact that languages which are similar genealogically (FAM) and geographically (GEO) have more common lexical stems. It means that Multi-SentAugment extracts sentences with lexically similar words which unlock the generalisation abilities of the model.

Does LayerAgg Enrich Semantic Content? While the task results seem to suggest this, we design a probing experiment which aims to answer the following question: *Do the representations obtained with LayerAgg really capture more semantic information?* To this end, we first obtain representations of the 1,000 most frequent words (Conneau et al., 2018; Mehri and Eric, 2021) in MultiATIS++[7] in each sentence using a frozen mBERT task-tuned on English, with and without LayerAgg. We then aim to identify which word was encoded by training a simple linear classifier. The rationale is that by storing more lexical information in the representations, similar words will obtain similar representations: consequently, the classifier should more easily identify the correct word.

The micro-averaged F_1 scores are shown in Table 6. The same positive trend with large gains in the classification score is observed in all languages, confirming our hypothesis. We note that the large gains are reported not only for English (which was used for task fine-tuning), but also in other languages, suggesting the benefits of LayerAgg in boosting cross-lingual lexical capabilities of multilingual encoders in transfer scenarios.

Cross-lingual Similarity in LayerAgg. We now assess how LayerAgg captures cross-lingual representation similarity by comparing self-attention maps for different languages emerging from Transformer $\mathbb{T}$. We analyse the similarity of representations of the source language (en) with each target language in MultiATIS++ and xSID using linear Centered Kernel Alignment (l-CKA, Kornblith et al. 2019), a standard tool for such analyses in Transformer-based models (Conneau et al., 2020b; Glavaš and Vulić, 2021). Linear CKA is a representation similarity metric for representations obtained from neural networks. L-CKA is invariant to orthogonal transformation and isotopic scaling (Glavaš and Vulić, 2021). More formally, it is defined as follows:

$$CKA(X, Y) = \frac{||Y^T X||_F^2}{||X^T X||_F ||Y^T Y||_F}$$

where X, Y are input matrices.

We measure 1) cross-lingual correspondence for slots where l-CKA is computed between the representations of the same slot[8] in different languages; 2) the correlation between the l-CKA scores and transfer performance.

The l-CKA scores for MultiATIS++ in Figure 4 reveal high similarities between self-attention maps for similar languages. For instance, the scores are high between Romance languages in MultiATIS++ and Germanic languages in XSID. At the same time, the scores are low between ja and Romance languages and between tr and all other, non-Turkic languages. Spearman's ρ correlation scores between the l-CKA scores and zero-shot transfer performance are also very strong. For MultiATIS++, $\rho = 0.95$ (intent classification) and $\rho = 0.92$ (slot labelling), while for xSID: $\rho = 0.77$ (intent classification) and $\rho = 0.59$ (slot labelling).

Another Multilingual Sentence Encoder? Intuitively, the effectiveness of Multi-SentAugment depends on the underlying multilingual sentence encoder $\mathbb{F}$. We now analyse how much performance

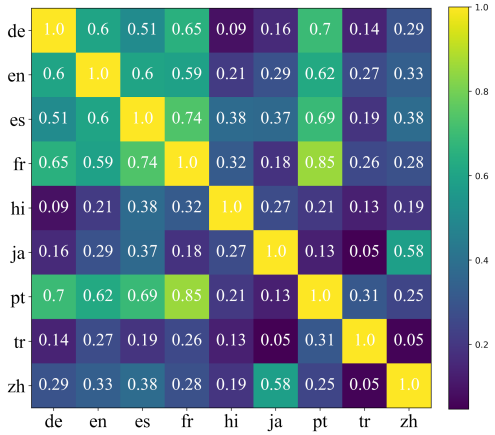


Figure 4: l-CKA similarities of mean-pooled representations of slots between different languages in MultiATIS++. For a similar plot for xSID see the Appendix.

[7] For a word tokenised into more than 1 WordPiece, we obtain its vector by averaging its constituent WordPiece vectors.

[8] Slot representation is the average of attention maps of tokens labelled with that slot. We cannot compare attention maps for each word/WordPiece directly: we lack alignments between the words across sentences in different languages.

Data size	𝔽	hi	ja	tr	AVG
Intent classification (Accuracy × 100)					
Few-shot	LASER	81.76	**79.28**	**78.87**	**79.97**
	LaBSE	**86.43**	77.16	69.36	77.65
Full-data	LASER	88.71	**96.42**	82.41	89.18
	LaBSE	**89.28**	**96.42**	**84.54**	**90.08**
Slot labelling (Slot F1 × 100)					
Few-shot	LASER	**73.34**	**81.92**	68.11	**74.46**
	LaBSE	69.88	81.19	**70.28**	73.78
Full-data	LASER	80.45	88.35	**73.32**	80.71
	LaBSE	**83.32**	**91.79**	71.86	**82.32**

Table 7: A comparison of LASER and LaBSE as underlying encoders for Multi-SentAugment. A model variant without LayerAgg used; very similar trends are observed with the *+LayerAgg* variant (see the Appendix).

differs if we replace one state-of-the-art encoder (i.e., LASER) with another: LaBSE (Feng et al., 2020), running Multi-SentAugment with LaBSE in 3 languages from 3 different language families that also use different scripts – Turkish, Hindi and Japanese. The results in Table 7 do indicate some performance variance across tasks and languages: LaBSE is slightly better in full-data scenarios while LASER performs better in few-shot scenarios. In future work on Multi-SentAugment, we will investigate encoder ensembles, and we plan to make the mining process more scalable and quicker.

7 Conclusion and Future Work

We presented **1)** LayerAgg, a layer aggregation method which learns to effectively combine useful semantic information from multiple layers of a pretrained multilingual Transformer, and **2)** Multi-SentAugment, a data augmentation approach that leverages unannotated Web-scale monolingual corpora to reduce manual annotation efforts. Our results suggest that both methods, applied with state-of-the-art multilingual dialogue NLU models, yield performance benefits both for intent classification and for slot labelling. The methods obtain consistent gains in zero-shot, few-shot and full-data setups on 2 multilingual NLU datasets spanning 16 languages. In future work, we will investigate further applications of Multi-SentAugment in cross-lingual settings (e.g., by mining sentences in languages from the same language family). We will also extend the methods towards truly low-resource languages. The code is available online at: `github.com/cambridgeltl/MultiSentAugment_LayerAgg`.

Acknowledgements

We thank the anonymous reviewers for their helpful comments and suggestions. This work is supported by the ERC PoC Grant MultiConvAI:: Enabling Multilingual Conversational AI (no. 957356), and a Huawei research donation.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Tanja Bunk, Daksh Varshneya, Vladimir Vlasov, and Alan Nichol. 2020. DIET: Lightweight language understanding for dialogue systems. *CoRR*, abs/2004.09936.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. 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, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

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.

Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. 2021. Self-training improves pre-training for natural language understanding. In *Proceedings of the 2021 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5408–5418, Online. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

Goran Glavaš and Ivan Vulić. 2021. Is supervised syntactic parsing beneficial for language understanding tasks? an empirical investigation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3090–3104, Online. Association for Computational Linguistics.

Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2020. Enriched pre-trained transformers for joint slot filling and intent detection. *arXiv preprint arXiv:2004.14848*.

Keqing He, Yuanmeng Yan, and Weiran Xu. 2020. Adversarial cross-lingual transfer learning for slot tagging of low-resource languages. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Prabhu Kaliamoorthi, Aditya Siddhant, Edward Li, and Melvin Johnson. 2021. Distilling large language models into tiny and effective students using pQRNN. *arXiv preprint arXiv:2101.08890*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR 2015*.

Dan Kondratyuk. 2019. Cross-lingual lemmatization and morphology tagging with two-stage multilingual BERT fine-tuning. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 12–18, Florence, Italy. Association for Computational Linguistics.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. 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 2779–2795.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR.

Jitin Krishnan, Antonios Anastasopoulos, Hemant Purohit, and Huzefa Rangwala. 2021. Multilingual code-switching for zero-shot cross-lingual intent prediction and slot filling. *arXiv preprint arXiv:2103.07792*.

Saurabh Kulshreshtha, Jose Luis Redondo Garcia, and Ching-Yun Chang. 2020. Cross-lingual alignment methods for multilingual BERT: A comparative study. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 933–942, Online. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2021. Evaluating multilingual text encoders for unsupervised cross-lingual retrieval. In *Advances in Information Retrieval*, pages 342–358. Springer International Publishing.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Liyuan Liu, Jingbo Shang, and Jiawei Han. 2019a. Arabic named entity recognition: What works and what's next. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 60–67, Florence, Italy. Association for Computational Linguistics.

Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 2019b. Zero-shot cross-lingual dialogue systems with transferable latent variables. 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 1297–1303, Hong Kong, China. Association for Computational Linguistics.

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2020. Zero-resource cross-domain named entity recognition. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 1–6, Online. Association for Computational Linguistics.

Maddalen López de Lacalle, Xabier Saralegi, and Iñaki San Vicente. 2020. Building a task-oriented dialog system for languages with no training data: the case for Basque. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2796–2802, Marseille, France. European Language Resources Association.

Samuel Louvan and Bernardo Magnini. 2020a. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 480–496, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Samuel Louvan and Bernardo Magnini. 2020b. Simple data augmentation for multilingual nlu in task oriented dialogue systems.

Samuel Louvan and Bernardo Magnini. 2020c. Simple is better! lightweight data augmentation for low resource slot filling and intent classification. In *33rd Pacific Asia Conference on Language, Information and Computation*.

Shikib Mehri and Mihail Eric. 2021. Example-driven intent prediction with observers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2979–2992, Online. Association for Computational Linguistics.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialisation of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the ACL*, 5:309–324.

Baolin Peng, Chenguang Zhu, Michael Zeng, and Jianfeng Gao. 2020. Data augmentation for spoken language understanding via pretrained models. *arXiv preprint arXiv:2004.13952*.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Libo Qin, Tianbao Xie, Wanxiang Che, and Ting Liu. 2021. A survey on spoken language understanding: Recent advances and new frontiers. *arXiv preprint arXiv:2103.03095*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Evgeniia Razumovskaia, Goran Glavaš, Olga Majewska, Anna Korhonen, and Ivan Vulić. 2021. Crossing the conversational chasm: A primer on multilingual task-oriented dialogue systems. *CoRR*, abs/2104.08570.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin

Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. 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.

Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Ari, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. 2020. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8854–8861.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Gokhan Tur and Renato De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

Gokhan Tur, Dilek Hakkani-Tür, and Larry Heck. 2010. What is left to be understood in atis? In *2010 IEEE Spoken Language Technology Workshop*, pages 19–24. IEEE.

Shyam Upadhyay, Manaal Faruqui, Gökhan Tür, Dilek Hakkani-Tür, and Larry P. Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *Proceedings of the 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.

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 *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Henry Weld, Xiaoqi Huang, Siqi Long, Josiah Poon, and Soyeon Caren Han. 2021. A survey of joint intent detection and slot-filling models in natural language understanding. *CoRR*, abs/2101.08091.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Lijun Wu, Yiren Wang, Yingce Xia, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019. Exploiting monolingual data at scale for neural machine translation. 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 4207–4216, Hong Kong, China. Association for Computational Linguistics.

Peng Xu, Hamidreza Saghir, Jin Sung Kang, Teng Long, Avishek Joey Bose, Yanshuai Cao, and Jackie Chi Kit Cheung. 2019. A cross-domain transferable neural coherence model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 678–687, Florence, Italy. Association for Computational Linguistics.

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.

Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, and Philip Yu. 2019. Joint slot filling and intent detection via capsule neural networks. In *Proceedings of*

the 57th Annual Meeting of the Association for Computational Linguistics, pages 5259–5267, Florence, Italy. Association for Computational Linguistics.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. A closer look at few-shot crosslingual transfer: The choice of shots matters. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5751–5767, Online. Association for Computational Linguistics.

Zijian Zhao, Su Zhu, and Kai Yu. 2019. Data augmentation with atomic templates for spoken language understanding. 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 3628–3634.

A Language Codes

en	English
ar	Arabic
da	Danish
de	German
de-st	South Tyrolean German dialect
es	Spanish
fr	French
hi	Hindi
id	Indonesian
it	Italian
ja	Japanese
kk	Kazakh
nl	Dutch
pt	Portuguese
sr	Serbian
tr	Turkish
zh	Chinese
th	Thai

Table 8: Language codes used in the paper.

B Training Hyperparameters

Hyperparameter	Value
Optimizer	Adam
Learning Rate	5e-5
Batch Size	32
BERT model	BERT base; multilingual cased
XLM-R model	XLM-R base

Table 9: Training hyperparameters.

C Full Results for Full-Data and 10-shot Setups

Target language	de	es	fr	hi	ja	pt	tr	zh	AVG
Intent classification (Accuracy × 100)									
Joint	96.08	95.07	**98.99**	79.13	78.12	90.59	73.19	**97.09**	88.53
+MSA	97.09	**96.86**	97.42	81.76	79.28	95.74	78.87	94.18	90.15
+MSA FILT	97.31	96.64	98.21	78.56	**82.53**	94.40	**79.15**	92.61	89.93
+LA	**98.10**	95.07	97.20	83.20	79.13	95.96	71.49	95.52	89.46
+LA +MSA	92.95	96.75	97.42	84.38	79.73	**96.87**	72.34	95.97	89.55
+LA +MSA FILT	97.87	91.94	97.47	**84.84**	79.73	95.63	78.87	96.08	**90.30**
Slot labelling (Slot F1 × 100)									
Joint	85.41	80.52	82.16	74.12	78.63	83.34	71.65	80.22	79.51
+MSA	82.95	80.70	82.41	73.34	81.92	84.10	68.11	**80.85**	79.30
+MSA FILT	**86.19**	81.90	82.79	76.02	**82.55**	83.62	66.82	76.18	79.51
+LA	85.50	82.13	82.62	73.80	75.64	84.37	71.92	73.40	78.67
+LA +MSA	85.48	**83.10**	**82.97**	72.87	80.99	84.46	68.46	76.30	79.33
+LA +MSA FILT	85.89	80.38	81.45	**76.71**	77.92	**85.00**	**74.24**	76.34	**79.74**

Table 10: Few-shot results on MultiATIS++. Acronyms: +MSA = +Multi-SentAugment; +MSA FILT = +Multi-SentAugment filtered by teacher model confidence; +LA = +LayerAgg; +LA +MSA = +LayerAgg +Multi-SentAugment; +LA +MSA FILT = +LayerAgg +Multi-SentAugment filtered by teacher model confidence. Highest scores in each task per column in **bold**. The underlying multilingual model is mBERT.

Target language	de	es	fr	hi	ja	pt	tr	zh	AVG
Intent classification (Accuracy × 100)									
Joint	98.65	**97.76**	97.87	88.26	95.97	97.98	84.26	94.66	94.43
+MSA	98.54	97.54	98.21	88.71	96.42	97.09	82.41	94.49	94.18
+MSA FILT	98.43	96.64	97.87	88.94	**96.75**	97.65	**85.82**	94.83	94.62
+LA	**98.88**	96.65	**98.54**	91.67	96.64	97.42	83.97	**96.98**	**95.09**
+LA +MSA	98.77	97.54	**98.54**	88.72	96.64	**98.10**	84.40	96.86	94.95
+LA +MSA FILT	98.66	97.31	97.65	**91.76**	**96.75**	97.42	82.84	**96.98**	94.92
Slot labelling (Slot F1 × 100)									
Joint	94.02	85.37	88.26	78.11	91.01	91.05	64.14	91.41	85.42
+MSA	94.02	85.05	**89.39**	80.45	88.35	91.06	**73.32**	90.93	86.57
+MSA FILT	93.65	85.12	88.77	80.78	90.56	90.99	67.41	91.67	86.12
+LA	**94.26**	85.73	89.02	80.92	92.03	90.77	71.09	92.33	**87.02**
+LA +MSA	93.16	85.69	89.10	**81.97**	**92.24**	**91.36**	70.14	91.59	86.91
+LA +MSA FILT	93.86	**85.96**	88.68	80.82	91.81	90.87	69.29	**92.52**	86.72

Table 11: Full-data results on MultiATIS++. Acronyms: +MSA = +Multi-SentAugment; +MSA FILT = +Multi-SentAugment filtered by teacher model confidence; +LA = +LayerAgg; +LA +MSA = +LayerAgg +Multi-SentAugment; +LA +MSA FILT = +LayerAgg +Multi-SentAugment filtered by teacher model confidence. Highest scores in each task per column in **bold**. The underlying multilingual model is mBERT.

D 5-shot and 20-shot Results with Multi-SentAugment

Target language	de	es	fr	hi	ja	pt	tr	zh	AVG
Intent classification (Accuracy × 100)									
Joint	96.19	94.63	96.08	63.74	78.28	95.07	60.00	**93.06**	84.63
+MSA	92.72	92.50	94.40	69.90	**81.64**	93.62	**64.26**	89.14	84.77
+MSA FILT	**97.20**	**96.87**	**97.31**	**77.77**	79.28	**95.19**	61.14	90.37	**86.89**
Slot labelling (Slot F1 × 100)									
Joint	**83.31**	77.66	**79.95**	67.00	72.32	82.5	62.66	**75.19**	75.08
+MSA	80.12	75.81	79.24	69.64	65.86	**82.72**	**62.81**	74.46	73.83
+MSA FILT	83.16	**79.25**	78.62	**70.49**	**74.30**	81.22	62.39	72.08	**75.19**

Table 12: 5-shot results of Multi-SentAugment on MultiATIS++. Acronyms: +MSA = +Multi-SentAugment; +MSA FILT = +Multi-SentAugment filtered by teacher model confidence. Highest scores in each task per column in **bold**. The underlying multilingual model is mBERT.

Target language	de	es	fr	hi	ja	pt	tr	zh	AVG
Intent classification (Accuracy × 100)									
Joint	97.54	89.81	97.65	84.38	**88.80**	92.05	77.30	87.46	89.37
+MSA	**97.65**	**95.97**	**98.43**	80.96	84.43	95.41	76.03	**93.62**	**90.31**
+MSA FILT	97.09	91.15	98.10	**87.57**	85.14	**96.53**	**78.30**	84.99	89.86
Slot labelling (Slot F1 × 100)									
Joint	88.93	**84.03**	**85.63**	73.15	82.12	85.09	**72.88**	78.05	81.24
+MSA	87.99	82.41	84.03	74.99	**82.38**	**85.37**	71.91	83.59	81.58
+MSA FILT	**88.94**	81.79	84.00	**76.56**	81.83	83.74	72.08	**84.13**	**81.63**

Table 13: 20-shot results of Multi-SentAugment on MultiATIS++. Acronyms: +MSA = +Multi-SentAugment; +MSA FILT = +Multi-SentAugment filtered by teacher model confidence. Highest scores in each task per column in **bold**. The underlying multilingual model is mBERT.

E Impact of Sentence Encoder (+LayerAgg Variant)

Model	F	hi	ja	tr	AVG
Intent classification (Acc times 100)					
Full-data	LASER	88.71	96.64	84.40	89.92
	LaBSE	90.08	96.98	83.55	**90.2**
Few-shot	LASER	84.28	79.73	72.34	**78.78**
	LaBSE	79.93	77.72	77.73	78.46
Slot labelling (Slot F1 times 100)					
Full-data	LASER	81.97	92.24	70.14	**81.45**
	LaBSE	82.85	91.40	69.62	**81.29**
Few-shot	LASER	72.87	80.99	68.46	**74.11**
	LaBSE	72.68	76.78	72.72	74.06

Table 14: Impact of the chosen multilingual sentence encoder: LASER (Artetxe and Schwenk, 2019) versus LaBSE (Feng et al., 2020) in full-data and few-shot scenarios for intent classification and slot labelling, for the LayerAgg model variant.

F l-CKA Similarities on xSID

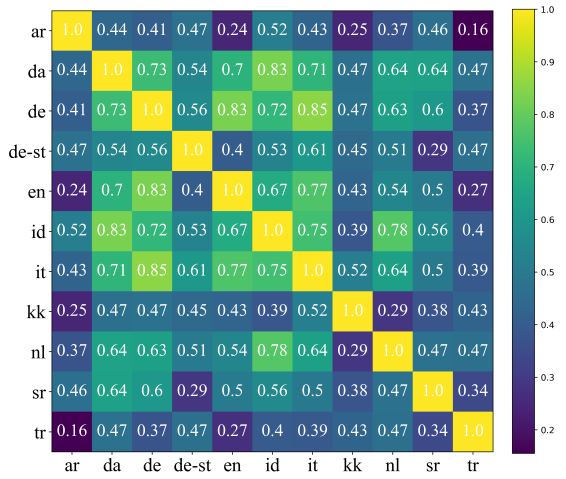


Figure 5: l-CKA similarities of mean-pooled representations of slots between different languages in xSID.