End-to-End Slot Alignment and Recognition for Cross-Lingual NLU

Weijia Xu[*]
University of Maryland
weijia@cs.umd.edu

Batool Haider
Amazon AI
bhaider@amazon.com

Saab Mansour
Amazon AI
saabm@amazon.com

Abstract

Natural language understanding (NLU) in the context of goal-oriented dialog systems typically includes intent classification and slot labeling tasks. Existing methods to expand an NLU system to new languages use machine translation with slot label projection from source to the translated utterances, and thus are sensitive to projection errors. In this work, we propose a novel end-to-end model that learns to align and predict target slot labels jointly for cross-lingual transfer. We introduce MultiATIS++, a new multilingual NLU corpus that extends the Multilingual ATIS corpus to nine languages across four language families, and evaluate our method using the corpus. Results show that our method outperforms a simple label projection method using fast-align on most languages, and achieves competitive performance to the more complex, state-of-the-art projection method with only half of the training time. We release our MultiATIS++ corpus to the community to continue future research on cross-lingual NLU.

1 Introduction

As a crucial component of goal oriented dialogue systems, natural language understanding (NLU) is responsible for parsing an utterance into a semantic frame to identify the user's need. These semantic frames are structured by what the user intends to do (the *intent*) and the arguments of the intent (the *slots*) (Tur et al., 2010). Given the English example in Figure 1, we identify the intent of the utterance as "flight" and label the slots to extract the departure city and airline name. Intent detection can be modeled as a sentence classification task where each utterance is labeled with an intent y_I. Slot filling is typically modeled as a sequence labeling task where given the utterance $\boldsymbol{x}_{1...n}$, each word x_i is labeled with a slot y_i.

Despite the high accuracy achieved by neural models on intent detection and slot filling (Goo et al., 2018; Qin et al., 2019), training such models on a new language requires additional efforts to collect large amounts of training data. One would consider transfer learning from high-resource to low-resource languages to minimize the efforts of data collection and annotation. However, currently available multilingual NLU datasets (Upadhyay et al., 2018; Schuster et al., 2019) only support three languages distributed in two language families, which hinders the study of cross-lingual transfer across a broad spectrum of language distances. In this paper, we release a new multilingual NLU corpus that contains training, development, and test data for six new languages in addition to the three languages in the Multilingual ATIS corpus (Upadhyay et al., 2018). The resulting corpus, namely MultiATIS++, consists in total of 37,084 training examples and 7,859 test examples covering nine languages in four language families.

Using our corpus, we explore the use of multilingual BERT encoder (Devlin et al., 2019), machine translation (MT), and label projection methods for multilingual training and cross-lingual transfer. Furthermore, we propose an end-to-end model for joint slot label alignment and recognition, so that it no longer relies on slot label projections using external word alignment tools (Mayhew et al., 2017; Schuster et al., 2019) or engineered features (Ehrmann et al., 2011; Jain et al., 2019), which may not generalize well to low-resource languages. Our model performs soft label alignment using an attention module which is trained jointly with other model components on intent classification, slot filling, and an augmented reconstruction objective designed to improve the soft label alignment.

Experimental results show that our method uses the same amount of training time as a simple label projection approach using fast-align, while achiev-

---
[*]Work done at Amazon AI.

ing significantly higher slot F1 on most languages. Furthermore, our method achieves competitive performance to the more complex, state-of-the-art label projection method that uses linguistic features to improve projection quality, while using half of the training time. Finally, our results show different trends when comparing various cross-lingual transfer methods on different languages, which emphasizes the need to evaluate cross-lingual transfer methods on a diverse set of languages to fully illustrate the strengths and weaknesses of each method.

2 Related Work

Cross-lingual transfer learning has been studied on a variety of sequence tagging tasks including part-of-speech tagging (Yarowsky et al., 2001; Täckström et al., 2013; Plank and Agić, 2018), named entity recognition (Zirikly and Hagiwara, 2015; Tsai et al., 2016; Xie et al., 2018), and natural language understanding (He et al., 2013; Upadhyay et al., 2018; Schuster et al., 2019). Existing methods can be roughly categorized into two categories: transfer through multilingual models and transfer through machine translation.

Transfer via Multilingual Models For closely related languages with similar alphabets, it is beneficial to train a multilingual model with shared character encoder to learn common character-based features (Yang et al., 2017; Lin et al., 2018). However, such techniques are less effective when applied to dissimilar languages that lack common lexical features. Chen et al. (2018, 2019) focus on the multi-source transfer scenario and apply adversarial training to extract language-invariant features shared by source languages. Recent advances on cross-lingual representations have enabled transfer between dissimilar languages. Representations from multilingual neural machine translation (NMT) encoders have been shown to be effective for cross-lingual text classification (Eriguchi et al., 2018; Yu et al., 2018; Singla et al., 2018). In this work, we use multilingual BERT (Devlin et al., 2019), an unsupervised cross-lingual language model trained on monolingual texts from a wide range of languages and has been shown to provide powerful sentence representations that lead to promising performance for zero-resource cross-lingual language understanding tasks (Pires et al., 2019). We leave the use of other recently proposed cross-lingual language models such as XLM-R (Conneau et al., 2020) to future work.

Transfer via Machine Translation requires translating the source language training data into the target language or translating the target language test data into the source language. Despite its empirical success on cross-lingual text classification tasks (Wan, 2009), it faces a challenging problem on the sequence tagging tasks: labels on the source language sentences need to be projected to the translated sentences. Most of the prior work relies on unsupervised word alignment from statistical MT (Yarowsky et al., 2001; Shah et al., 2010; Ni et al., 2017) or attention weights from NMT models (Schuster et al., 2019). Other heuristic approaches include matching tokens based on their surface forms (Feng et al., 2004; Samy et al., 2005; Ehrmann et al., 2011) and more complex projection approaches that combine linguistic features with information from the MT systems (Jain et al., 2019). By contrast, our method does not rely on external word alignment or linguistic features, but models label projection through an attention module that can be jointly trained with other model components on the machine translated data.

3 Data

One of the most popular datasets for multilingual NLU is the ATIS dataset (Price, 1990) and its multilingual extension (Upadhyay et al., 2018). The ATIS dataset is created by asking each participant to interact with an agent (who has access to a database) to solve a given air travel planning problem. Upadhyay et al. (2018) extend the English ATIS to Hindi and Turkish by manually translating and annotating a subset of the training and test data via crowdsourcing.[1]

To facilitate study on cross-lingual transfer across a broader spectrum of language distances, we create the MultiATIS++ corpus by extending both the training and test set of the English ATIS corpus to six additional languages.[2] The resulting corpus covers nine languages in four different language families including Indo-European (English, Spanish, German, French, Portuguese, and Hindi), Sino-Tibetan (Chinese), Japonic (Japanese), and Altaic (Turkish).

For each new language, we hire professional native translators to translate the English utterances and annotate the slots at the same time. When

---

[1] https://catalog.ldc.upenn.edu/LDC2019T04
[2] https://github.com/amazon-research/multiatis

EN	show	departures	from	atlanta	for	american
	O	O	O	B-fromloc.city_name	O	B-airline_name
ES	Muestra	salidas	desde	Atlanta	de	American
	O	O	O	B-fromloc.city_name	O	B-airline_name
PT	Mostre	partidas	de	Atlanta	da	American
	O	O	O	B-fromloc.city_name	O	B-airline_name
DE	Zeige	Abflüge	von	Atlanta	für	American
	O	O	O	B-fromloc.city_name	O	B-airline_name
FR	Montrer	des	départs	d'	Atlanta	pour	American
	O	O	O	O	B-fromloc.city_name	O	B-airline_name
ZH	显示	美国航空	从	亚特兰大	出发的航班		
	O	B-airline_name	O	B-fromloc.city_name	O		
JA	アトランタ	発	アメリカ	便を表示する			
	B-fromloc.city_name	O	B-airline_name	O			
HI	अमेरिकन	के	लिए	अटलांटा	से	प्रस्थान दिखाएं	
	B-airline_name	O	O	B-fromloc.city_name	O	O O	
TR	atlanta	'	dan	american	kalkislarini	goster	
	B-fromloc.city_name	O	O	B-airline_name	O	O	

Figure 1: An English training example and its translated versions in the MultiATIS++ corpus. The English utterance is manually translated to the other eight languages including Spanish (ES), Portuguese (PT), German (DE), French (FR), Chinese (ZH), Japanese (JA), Hindi (HI), and Turkish (TR). For each language, we show the utterance followed by the slot labels in the BIO format. The intent of the utterances is the *flight* intent.

Language	Utterances			Tokens			Intents	Slots
	train	dev	test	train	dev	test		
English	4488	490	893	50755	5445	9164	18	84
Spanish	4488	490	893	55197	5927	10338	18	84
Portuguese	4488	490	893	55052	5909	10228	18	84
German	4488	490	893	51111	5517	9383	18	84
French	4488	490	893	55909	5769	10511	18	84
Chinese	4488	490	893	88194	9652	16710	18	84
Japanese	4488	490	893	133890	14416	25939	18	84
Hindi	1440	160	893	16422	1753	9755	17	75
Turkish	578	60	715	6132	686	7683	17	71

Table 1: Data statistics for the MultiATIS++ corpus. The number of *utterances* and *tokens* (characters for Chinese and Japanese) are provided for the training (train), development (dev), and test sets for each of the nine languages. The total number of *intents* and *slots* (before adding the BIO tags) are also given.

translating, we ask the translators to preserve the spoken modality phenomena (e.g. hesitations and word repetitions) and style (e.g. degree of formality) of the original English sentences, so that it is closer to the real scenaria. To get the slot labels for the translated utterances, we ask the translators to tag the segments in the translated utterances that are aligned to the corresponding English segments. Finally, we tokenize the translated utterances and BIO tag the tokens. For quality control, we ask a third-party to perform several rounds of qualification checks until no issues are reported.

We show an English training example and its translated versions in the other eight languages in Figure 1 and report the data statistics in Table 1. We split the English training and development sets randomly and keep the same split for all the other

languages except for Hindi and Turkish from the Multilingual ATIS corpus. Note that the Hindi and Turkish portions of the data are smaller than the other languages, covering only a subset of the intent and slot types.

4 Cross-Lingual NLU

4.1 Joint Intent Detection and Slot Filling

Following Liu and Lane (2016), we model intent detection and slot filling jointly. We add a special classification token x_0 at the beginning of the input sequence $\boldsymbol{x} = (x_1, x_2, ..., x_T)$ of length T following Devlin et al. (2019). Next, an encoder Θ_{enc} is used to produce a sequence of contextualized representations $\boldsymbol{h}_{0...T}$ given the input sequence

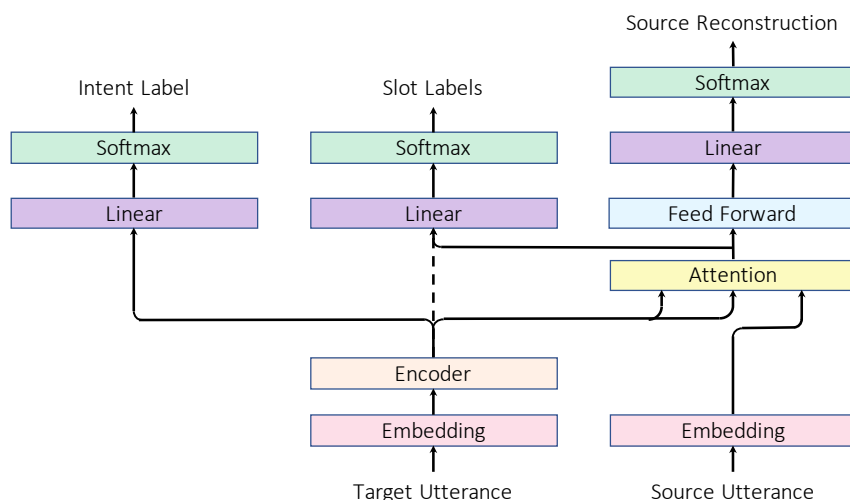$$\boldsymbol{h}_{0...T} = \Theta_{enc}(x_0, x_1, ..., x_T)$$

Figure 2: Architecture of our soft-alignment model for end-to-end slot alignment and recognition. The model is trained without external label projection: it learns to soft-align the representations of the target utterance to the source slot labels and predict the intent and slot labels jointly. The dotted line denotes the path during inference, where we directly connect the encoder module to the intent and slot classification layer to make predictions on the target utterance.

For intent detection, we take the representation $\boldsymbol{h}_0$ corresponding to x_0 as the sequence representation and apply a linear transformation and a softmax function to predict the intent probability

$$p_{\text{intent}}(\cdot|\boldsymbol{x}) = \text{softmax}(\boldsymbol{W}^I \boldsymbol{h}_0 + \boldsymbol{b}^I)$$

For slot filling, we compute the probability for each slot using the representations $\boldsymbol{h}_{1\ldots T}$

$$p_{\text{slot}_i}(\cdot|\boldsymbol{x}) = \text{softmax}(\boldsymbol{W}^S \boldsymbol{h}_i + \boldsymbol{b}^S)$$

We explore two different encoder models:

- **LSTM:** We use the concatenation of the forward and backward hidden states of a bidirectional LSTM (Schuster and Paliwal, 1997) as the encoder representations. We initialize the encoder and embeddings randomly.

- **Multilingual BERT:** We use the multilingual BERT (Devlin et al., 2019) pre-trained in an unsupervised way on the concatenation of monolingual corpora from 104 languages. We take the hidden states from the top layer as the encoder representations and fine-tune the model on the NLU data.

4.2 Problems in Slot Label Projection

Past work has shown the effectiveness of using MT systems to boost the performance of cross-lingual NLU (Schuster et al., 2019). More specifically, one first translates the English training data to the target

language using an MT system, and then projects the slot labels from English to the target language. Prior work projects the slot labels using word alignments from statistical MT models (Yarowsky et al., 2001) or attention weights from neural MT models (Schuster et al., 2019). The final performance on the target language highly depends on the quality of the slot projection. Jain et al. (2019) show that improving the quality of projection leads to significant improvements in the target performance on cross-lingual named entity recognition. However, the improvements come at the cost of much more complex and expensive projection process using engineered features.

To address the above issues, we propose a soft-alignment model that performs end-to-end slot alignment and recognition using an additional attention module (Figure 2), so that it requires no external slot projection process. Furthermore, we show that the soft slot alignment can be strengthened by building it on top of strong encoder representations from multilingual BERT.

4.3 End-to-End Slot Alignment and Recognition via Attention

Given a source utterance $\boldsymbol{s}_{1\ldots S}$ of length S and its translation $\boldsymbol{t}_{1\ldots T}$ of length T in the target language, the model learns to predict the target slot labels and soft-align it with the source labels via attention. First, it encodes the source utterance into a sequence of embeddings $\boldsymbol{e}_{1\ldots S}^{(src)}$ and then encodes

the translation $\boldsymbol{t}_{0...T}$ (t_0 is inserted as the classification token) into a sequence of contextualized representations $\boldsymbol{h}_{0...T}^{(tgt)} = \Theta_{enc}(\boldsymbol{t}_{0...T})$, where Θ_{enc} represents the encoder. For intent classification, we assume that the translated utterance has the same intent as the source utterance. Thus we compute the intent probabilities using the representation $\boldsymbol{h}_0^{(tgt)}$

$$p_{\text{intent}}(\cdot|\boldsymbol{t}) = \text{softmax}(\boldsymbol{W}^I \boldsymbol{h}_0^{(tgt)} + \boldsymbol{b}^I)$$

and the intent classification loss given the intent label $y_I^{(src)}$ on the source utterance

$$\mathcal{L}_{\text{intent}} = -\log p_{\text{intent}}(y_I^{(src)}|\boldsymbol{t})$$

For slot filling, we introduce an attention module to connect the source slot labels $\boldsymbol{y}_{1...S}^{(src)}$ with the target sequence $\boldsymbol{t}_{1...T}$. First, we compute the hidden state at each source position as a weighted average of the target representations

$$\boldsymbol{z}_i = \sum_{j=1}^{T} a_{ij} \boldsymbol{h}_j^{(tgt)}$$

where $\boldsymbol{z}_i$ is the hidden state at source position i, and a_{ij} is the attention weights between the source word s_i and translation word t_j. To compute the weights a_{ij}, we first linearly project the query vector $\boldsymbol{e}_i^{(src)}$ and the key vector $\boldsymbol{h}_j^{(tgt)}$ with learnable parameters to d dimensions. We then perform the scaled dot-product attention on the projected query and key vectors

$$\boldsymbol{a}_i = \text{softmax}\left(\frac{(\boldsymbol{e}_i^{(src)}\boldsymbol{W}^Q)(\boldsymbol{h}^{(tgt)}\boldsymbol{W}^K)^{\mathsf{T}}}{\sqrt{d}\tau}\right)$$

where the projections $\boldsymbol{W}^Q$ and $\boldsymbol{W}^K$ are parameter matrices, and τ is a hyperparameter that controls the temperature of the softmax function.

Next, we compute the slot probabilities at the source position i using the hidden state $\boldsymbol{z}_i$

$$p_{\text{slot}_i}(\cdot|s_i, \boldsymbol{t}) = \text{softmax}(\boldsymbol{W}^S \boldsymbol{z}_i + \boldsymbol{b}^S)$$

and the slot filling loss given the slot labels $y_{1...S}^{(src)}$ on the source utterance

$$\mathcal{L}_{\text{slot}} = -\sum_{i=1}^{S} \log p_{\text{slot}_i}(y_i^{(src)}|s_i, \boldsymbol{t})$$

In addition, to improve the attention module to better align the source and target utterances, we add a reconstruction module consisting of a position-wise feed-forward and a linear output layer[3] to recover the source utterance using the attention outputs. We compute the probability distribution over the source vocabulary at position i as

$$p_{\text{rec}_i}(\cdot|s_i, \boldsymbol{t}) = \text{softmax}(\boldsymbol{W}^R \tilde{\boldsymbol{z}}_i + \boldsymbol{b}^R)$$
$$\tilde{\boldsymbol{z}}_i = \text{FeedForward}(\boldsymbol{z}_i) \tag{1}$$

and the reconstruction loss as

$$\mathcal{L}_{\text{rec}} = -\sum_{i=1}^{S} \log p_{\text{rec}_i}(s_i|s_i, \boldsymbol{t}) \tag{2}$$

The final training loss is $\mathcal{L} = \mathcal{L}_{\text{intent}} + \mathcal{L}_{\text{slot}} + \mathcal{L}_{\text{rec}}$.

Empirically, we find it beneficial to train the model jointly on the machine translated target data using the objective $\mathcal{L}$ and the source data using the supervised objective.

The attention and reconstruction modules are only used during training. During inference, we directly feed the encoder representations $\boldsymbol{h}_{0...T}^{(tgt)}$ of the target language utterance to the intent and slot classification layers

$$p_{\text{intent}}(\cdot|\boldsymbol{t}) = \text{softmax}(\boldsymbol{W}^I \boldsymbol{h}_0^{(tgt)} + \boldsymbol{b}^I)$$
$$p_{\text{slot}_i}(\cdot|\boldsymbol{t}) = \text{softmax}(\boldsymbol{W}^S \boldsymbol{h}_i^{(tgt)} + \boldsymbol{b}^S) \tag{3}$$

5 Multilingual NLU

In our first set of experiments, we explore using pre-trained multilingual BERT encoder for multilingual NLU. We compare the following training strategies to leverage the full supervised training data:

- **Target only:** Train each model on the target language training data using the standard supervised objective.

- **Multilingual:** Train a model on the concatenation of training data from all languages using the standard supervised objective.

Setup We train the models using the Adam optimizer (Kingma and Ba, 2015) for 20 epochs and select the model that performs the best on the development set (details in Appendix A). Following (Goo et al., 2018), we use intent accuracy and slot F1 as evaluation metrics.

---

[3]We tie the output weights $\boldsymbol{W}^R$ with BERT embeddings.

Intent acc.		en	es	de	zh	ja	pt	fr	hi	tr
Target only	LSTM	96.08	93.04	94.02	92.50	91.18	92.70	94.71	84.46	81.12
	BERT	**97.20**	**96.44**	**96.73**	**95.52**	95.54	**96.71**	**97.38**	90.50	87.10
Multilingual	LSTM	95.45	94.09	95.05	93.42	92.90	94.02	94.80	87.79	85.43
	BERT	**97.20**	**96.77**	**96.86**	**95.54**	**96.44**	96.48	**97.24**	**92.70**	**92.20**
Slot F1		en	es	de	zh	ja	pt	fr	hi	tr
Target only	LSTM	94.71	75.89	91.44	90.84	88.80	88.43	85.93	74.93	64.43
	BERT	95.57	86.58	**94.98**	**93.52**	91.40	91.35	89.14	82.36	75.21
Multilingual	LSTM	94.75	84.11	92.00	90.76	88.55	88.79	87.96	77.34	77.25
	BERT	**95.90**	**87.95**	**95.00**	**93.67**	**92.04**	**91.96**	**90.39**	**86.73**	**86.04**

Table 2: Results on MultiATIS++ using full training data and the standard supervised objective averaged over 5 runs. The *Target only* models are trained only on the target language training data. The *Multilingual* models are trained on the concatenation of training data from all languages.

Results Table 2 shows the results using full training data and the supervised objective. First, we compare LSTM and BERT models trained on the target language data only. Multilingual BERT encoder brings significant[4] improvements of 1–6% on intent accuracy and 1–11% on slot F1. The largest improvements are on the two low-resource languages: Hindi and Turkish. Multilingual training on all languages brings further improvements on Hindi and Turkish: it improves intent accuracy by 2–5% and slot F1 by 4–11% for both LSTM and BERT models.

Comparison with SOTA On English ATIS, Qin et al. (2019) report 97.5% intent accuracy and 96.1% slot F1 when using BERT with their proposed stack-propagation architecture. This is comparable to our *target only with BERT* scores in Table 2. On multilingual ATIS, Upadhyay et al. (2018) report slot F1 of 80.6% on Hindi and 78.9% on Turkish using bilingual training. Our multilingual BERT model achieves higher F1 by +6.1% on Hindi and +7.1% on Turkish.

6 Cross-Lingual Transfer

In this section, we compare the following methods for cross-lingual transfer where we only use the English training data and a small amount (few-shot) or no (zero-shot) training data from the target language:

- **No MT:** Train the models only on the English training data without machine translating them to the target language.

- **MT+fast-align:** Use MT systems to translate the English data to the target language and project the slot labels using word alignment from *fast-align*.[5]

- **MT+TMP:** Use MT to translate the English data to the target language and the Translate-Match-Project method (Jain et al., 2019) to project the slot labels.

- **MT+soft-align:** Use MT to translate the English data to the target language and our soft-alignment method described in Section 4.3.

For all three MT-based approaches, we use AWS Translate for automatic translation and perform multilingual training on the English and machine translated data. We adopt the same setup as the previous section, except that we select the model at the last epoch as we assume no access to the development set from the target language in this setting. For the attention module in our soft-alignment model, we set the temperature $\tau = 0.1$.

6.1 Zero-Shot Results

Table 3 shows the results on zero-shot cross-lingual transfer. First, we study the impact of using multilingual BERT for MT-based transfer approaches. For *MT+fast-align*, BERT boosts the performance over LSTM by large margins: 14–32% on intent accuracy and 29–61% on slot F1 on all languages except for Turkish – a dissimilar language to English. For both LSTM and BERT models, using *MT+fast-align* brings significant improvements over their counterparts without MT on intent

Intent acc.		es	de	zh	ja	pt	fr	hi	tr
No MT	LSTM	64.82	64.77	59.69	65.40	65.15	69.92	60.11	63.64
	BERT	96.35	95.27	86.27	79.42	94.96	95.92	80.96	69.59
MT+fast-align	LSTM	95.36	94.02	92.70	77.96	94.20	94.76	88.71	87.72
	BERT	**97.02**	**96.77**	**96.10**	**88.82**	96.55	96.89	**93.12**	**93.77**
MT+TMP	LSTM	95.32	93.03	90.41	84.43	94.36	94.15	88.67	87.47
	BERT	**97.00**	96.01	95.16	**88.51**	96.46	97.04	92.41	**93.74**
MT+soft-align	BERT	97.20	96.66	95.99	88.33	96.78	97.49	92.81	93.71
Slot F1		es	de	zh	ja	pt	fr	hi	tr
No MT	LSTM	27.98	32.96	1.60	2.71	25.52	29.70	2.26	26.56
	BERT	74.98	82.61	62.27	35.75	74.05	75.71	31.21	23.75
MT+fast-align	LSTM	76.30	83.83	78.61	70.23	76.28	64.37	60.02	21.53
	BERT	79.18	87.21	81.82	79.53	78.26	70.18	69.42	23.61
MT+TMP	LSTM	79.19	84.99	82.84	71.98	79.44	77.52	67.70	40.14
	BERT	**83.98**	87.54	**85.05**	**82.60**	**81.73**	**79.80**	77.24	44.80
MT+soft-align	BERT	76.42	**89.00**	83.25	79.10	76.30	**79.64**	**78.56**	**61.70**

Table 3: Zero-shot results on MultiATIS++ averaged over 5 runs. The *No MT* rows are models trained only on the English data. The *MT+fast-align* rows correspond to models trained on the English and machine translated data with automatically projected slot labels using *fast-align*, and *MT+TMP* correspond to Translate-Match-Project (Jain et al., 2019). The *MT+soft-align* row is the model trained on the English and machine translated data using our soft-alignment method.

		Time (mins)
No MT	LSTM	18
	BERT	65
MT+fast-align	LSTM	232
	BERT	317
MT+TMP	LSTM	626
	BERT	719
MT+soft-align	BERT	352

Table 4: Total training time of each method on all languages for the zero-shot transfer experiments (including the time for machine translation, label projection, and model training). *MT+TMP* requires a time-consuming label projection process and thus takes double the time as our soft-alignment method.

accuracy – improvements of 13–33% when using LSTM and 1–24% when using BERT. However, we observe different trends on slot F1 for different languages. For example, when using BERT, *MT+fast-align* improves slot F1 by 20–44% on Chinese, Japanese, and Hindi over BERT without MT, but hurts by around 6% on French.[6] This is possibly because that the mBERT representations of French are of high quality, which leads to relatively high slot F1 without MT, thus adding more noisy data

---

via *MT+fast-align* does more harm than good. This indicates that, while training directly on the target language data is beneficial especially for languages dissimilar to English, the noisy projection of the slot labels could also bring harm to the model.

Next, we compare our soft-alignment method with other MT-based approaches. Our method outperforms *MT+fast-align* on five of the eight languages and is more robust across languages – it achieves consistent improvements over BERT fine-tuned without MT on both intent accuracy and slot F1, while *MT+fast-align* leads to a degradation on French and Turkish. Furthermore, we compare our method with *MT+TMP*, a strong label projection baseline that combines MT and linguistic features to improve the quality of label projection. Results show that our method uses only half the training time as *MT+TMP* (Table 4), while achieving competitive performance (Table 3): on intent accuracy, our method achieves on par or higher scores on all languages. On slot F1, our method performs on par or better than *MT+TMP* on four of the eight languages and achieves close performance on the remaining languages.

Finally, these results emphasize the need to evaluate cross-lingual transfer methods on a diverse set of languages to fully illustrate the strengths and
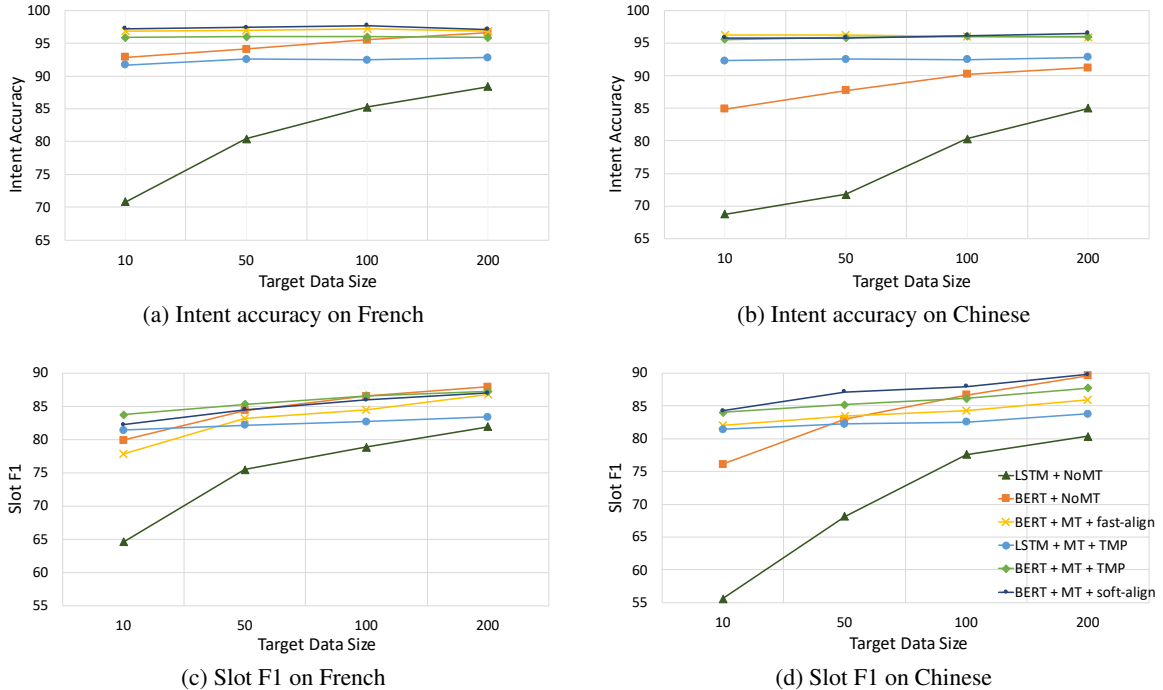
Figure 3: Results for cross-lingual transfer given various sizes of the labeled data from the target language in addition to the English and machine translated data used in zero-shot experiments. We report scores averaged over 5 runs using target training data selected with different random seeds. Results show the effectiveness of our soft-alignment method as compared to the best projection-based method (*BERT+MT+TMP*) in few-shot transfer where only a few hundred target examples are available.

weaknesses of each method and to avoid overly strong claims. In our experiments, we find that although *MT+TMP* achieves higher slot F1 than our method on some languages, our method is more robust across languages and outperforms *MT+TMP* by a large margin (+16.9%) on Turkish. To further explain the large gap, we measure the slot projection accuracy of *fast-align* and *TMP* compared against the annotated data. We find that *fast-align* obtains extremely low projection accuracy (20%) on Turkish, and *TMP* improves it to 39%, which is still low compared to that on other languages (above 70%). The low projection accuracy on Turkish can be attributed to the morphological difference between Turkish and English, which makes it difficult to hard-align each Turkish word to a single English word.

6.2 Learning Curves

Figure 3 shows the few-shot transfer results where we add a small amount of labeled data from the target language.[7] We select French as a similar language to English, and Chinese as a dissimilar

[7] We apply only the standard supervised objective on the target language data.

Intent acc.	average
MT+soft-align	94.87
w/o reconstruction loss	94.95
w/o joint training on source	94.64
Slot F1	**average**
MT+soft-align	80.00
w/o reconstruction loss	76.42
w/o joint training on source	71.62

Table 5: Ablation results for zero-shot transfer learning. Scores are averaged over Spanish, German, Chinese, Japanese, Portuguese, French, Hindi, and Turkish. Ablating the reconstruction loss or joint training on the source language data using supervised objective leads to a major drop on slot F1, while the impact on intent accuracy is small.

language. We find that *BERT+NoMT* obtains promising results in the few-shot setting with only several hundred training examples in the target language – it achieves comparable or even higher scores than the best MT-based approach on French, but still lags behind on Chinese by around 5% on intent accuracy. In addition, results

show the effectiveness of our method in few-shot transfer: our method obtains comparable (on French) or higher (on Chinese) slot F1 than the best projection-based method (*BERT+MT+TMP*) given a few hundred target examples, which suggests that, even with a less noisy label projector, the projection errors may still hinder the model from best exploiting the small amount of target language data especially on languages dissimilar to English.

6.3 Ablation Study

We evaluate the impact of different components in our soft-alignment model. Table 5 shows that both the reconstruction loss and joint training on the source data using supervised objective are beneficial – slot F1 drops by 3.6% when ablating the reconstruction loss and by 8.4% when ablating joint training on the source, while both have little impact on intent accuracy.

7 Conclusion

We introduce MultiATIS++, a multilingual NLU corpus that extends the Multilingual ATIS corpus to nine languages across four language families. We use our corpus to evaluate various cross-lingual transfer methods including the use of multilingual BERT encoder, machine translation, and label projection. We further introduce a novel end-to-end model for joint slot label alignment and recognition that requires no external label projection. Experiments show that multilingual BERT brings substantial improvements on multilingual training and cross-lingual transfer tasks. Furthermore, our model outperforms the simple projection baseline using fast-align on most languages, and achieves competitive performance to the state-of-the-art label projection approach with only half of the training time. We release our MultiATIS++ corpus to facilitate future research on cross-lingual NLU to bridge the gap between cross-lingual transfer and supervised methods.

Acknowledgments

References

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112. Association for Computational Linguistics.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

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.

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. Association for Computational Linguistics.

Maud Ehrmann, Marco Turchi, and Ralf Steinberger. 2011. Building a multilingual named entity-annotated corpus using annotation projection. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 118–124. Association for Computational Linguistics.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *CoRR*, abs/1809.04686.

Donghui Feng, Yajuan Lv, and Ming Zhou. 2004. A new approach for English-Chinese named entity alignment. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 372–379. Association for Computational Linguistics.

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 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757. Association for Computational Linguistics.

Xiaodong He, Li Deng, Dilek Hakkani-Tur, and Gokhan Tur. 2013. Multi-style adaptive training for robust cross-lingual spoken language understanding. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8342–8346. IEEE.

Alankar Jain, Bhargavi Paranjape, and Zachary C. Lipton. 2019. Entity projection via machine translation for cross-lingual NER. 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 1083–1092. Association for Computational Linguistics.

Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Ying Lin, Shengqi Yang, Veselin Stoyanov, and Heng Ji. 2018. A multi-lingual multi-task architecture for low-resource sequence labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 799–809. Association for Computational Linguistics.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *Interspeech 2016*, pages 685–689.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545. Association for Computational Linguistics.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Barbara Plank and Željko Agić. 2018. Distant supervision from disparate sources for low-resource part-of-speech tagging. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 614–620. Association for Computational Linguistics.

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*.

Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu. 2019. A stack-propagation framework with token-level intent detection 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 2078–2087, Hong Kong, China. Association for Computational Linguistics.

Doaa Samy, Antonio Moreno, and Jose M Guirao. 2005. A proposal for an arabic named entity tagger leveraging a parallel corpus. In *International Conference RANLP, Borovets, Bulgaria*, pages 459–465.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

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. Association for Computational Linguistics.

Rushin Shah, Bo Lin, Anatole Gershman, and Robert Frederking. 2010. Synergy: a named entity recognition system for resource-scarce languages such as swahili using online machine translation. In *Proceedings of the Second Workshop on African Language Technology (AfLaT 2010)*, pages 21–26.

Karan Singla, Dogan Can, and Shrikanth Narayanan. 2018. A multi-task approach to learning multilingual representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 214–220. Association for Computational Linguistics.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. 2016. Cross-lingual named entity recognition via wikification. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 219–228. Association for Computational Linguistics.

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, Gokhan Tur, Dilek Hakkani-Tur, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *Proceedings of the IEEE ICASSP*.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 235–243. Association for Computational Linguistics.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379. Association for Computational Linguistics.

Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. In *International Conference on Learning Representations*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.

Katherine Yu, Haoran Li, and Barlas Oguz. 2018. Multilingual seq2seq training with similarity loss for cross-lingual document classification. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 175–179. Association for Computational Linguistics.

Ayah Zirikly and Masato Hagiwara. 2015. Cross-lingual transfer of named entity recognizers without parallel corpora. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 390–396. Association for Computational Linguistics.

A   Model and Training Details

We train all models on 4 NVIDIA V100 Tensor Core GPUs. For both the multilingual NLU and cross-lingual transfer experiments, we train the models using the Adam optimizer (Kingma and Ba, 2015) for 20 epochs. We set the initial learning rate to 10^{-3} for the LSTM model and 10^{-5} for the BERT model. The LSTM model has embeddings of size 256 and 128 hidden units. We add dropout of 0.1 to the embeddings and encoder hidden states. Both LSTM and BERT models use the WordPiece tokenization model from (Devlin et al., 2019). Table 6 shows the total number of parameters for each model, and Table 7 shows the training time used for each method in the multilingual NLU experiments.

	#params
LSTM	$27,539,866$
BERT	$166,884,250$
BERT+soft-align	$169,944,625$

Table 6: Total number of parameters for each model. Our model contains additional parameters for the attention module and linear output layer for the reconstruction loss.

		Time (mins)
Target only	LSTM	18
	BERT	65
Multilingual	LSTM	9
	BERT	51

Table 7: Total training time for all languages in the multilingual NLU experiments.

B   Evaluation

We evaluate all NLU models using intent accuracy and slot F1. Before computing slot F1, we merge all slots that are segmented during preprocessing to match with the original slot segments. We use the script `conlleval.pl`[8] to compute slot F1.

C   Validation Performance

We report the average intent accuracy and slot F1 on the development sets in the multilingual NLU

---

experiments in Table 8. For zero-shot experiments, we select the model at the last epoch as we assume no access to the development sets from the target language in this setting.

Intent acc.		Average
Target only	LSTM	94.98
	BERT	97.12
Multilingual	LSTM	98.57
	BERT	98.37
Slot F1		Average
Target only	LSTM	89.41
	BERT	91.91
Multilingual	LSTM	97.64
	BERT	98.82

Table 8: Average intent accuracy and slot F1 on the development sets in the multilingual NLU experiments. For *targe only* models, we average the scores over all nine languages. For *multilingual* models, we only validate on the English development set.