E2E slot alignment and recognition for Cross-lingual NLU

The authors of this paper introduce an extended version of the dataset ATIS, which contains NLU flight data, with more languages. The new dataset is called MultiATIS++ and contains 9 different languages. In addition, the authors propose a new method for jointly aligning and predicting slots. This is supposed to solve the issues with slot projection that are dependent on external word alignment tools or complex heuristics.

In general there are two methods for cross-lingual transfer learning. The first is transfer via cross-lingual representations. An example for this is multilingual BERT, which was used in this paper. The second is transfer via machine translation. One of the challenges with this option is the need to project the labels from the source to the target language. The authors solve that issue through an attention mode that can be jointly trained with other model components on the MT data. They show that the soft-alignment method works better than the projection method. Also, that mBERT performs really well in all data scales.

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 in the context of goal oriented dialog systems typically includes intent classification and slot labeling tasks. An effective method to expand an NLU system to new languages is using machine translation (MT) with annotation projection to the target language. Previous work focused on using word alignment tools or complex heuristics for slot annotation projection. In this work, we propose a novel end-to-end model that learns to align and predict slots.

Existing multilingual NLU data sets only support up to three languages which limits the study on cross-lingual transfer. To this end, we construct a multilingual NLU corpus, MultiATIS++, by extending the Multilingual ATIS corpus to nine languages across various language families. We use the corpus to explore various cross-lingual transfer methods focusing on the zero-shot setting and leveraging MT for language expansion. Results show that our soft-alignment method significantly improves slot F1 over strong baselines on most languages. In addition, our experiments show the strength of using multilingual BERT for both cross-lingual training and zero-shot transfer.

1 Introduction

As a crucial component of goal oriented dialogue systems, natural language understanding (NLU) is responsible for parsing the user's 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 $x_{1...n}$, 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 data sets (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 introduce a multilingual NLU corpus by extending the Multilingual ATIS corpus (Upadhyay et al., 2018), an existing NLU corpus that includes training and test data for English, Hindi, and Turkish, with six new languages including Spanish, German, Chinese, Japanese, Portuguese, and French. 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 evaluate the recently proposed multilingual BERT encoder (Devlin et al., 2019) on the cross-lingual training and zero-shot transfer tasks. In addition, we identify a major drawback in the traditional transfer methods using machine translation (MT): they rely on slot label projections by external word alignment tools (Mayhew et al., 2017; Schuster et al., 2019) or complex heuristics (Ehrmann et al., 2011; Jain et al., 2019) which may not be generalizable to other tasks or lower-resource languages. To address the problem, we propose an end-to-end model that learns to jointly align and predict slots, so that the soft

---

[*] Work performed while interning at Amazon AI

slot alignment is improved jointly with other model components and can potentially benefit from powerful cross-lingual language encoders like multilingual BERT.

Experimental results show that our soft-alignment method achieves significantly higher slot F1 scores than the traditional projection method on most languages and leads to continuous improvements when the size of the target training data increases, while the traditional method quickly plateaus. In addition, our experiments show the strength of using multilingual BERT on both the cross-lingual training and zero-shot transfer tasks. When given a small amount of annotated data in the target language, multilingual BERT achieves comparable or even higher scores than the translation methods on slot F1.

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 cross-lingual representations and transfer through machine translation.

Transfer via Cross-Lingual Representations Recent advances on cross-lingual sequence encoders have enabled transfer between dissimilar languages. Representations learned by 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). However, these methods still rely on high-quality NMT encoders trained on large amounts of parallel data. In this work, we explore using multilingual BERT (Devlin et al., 2019), a cross-lingual language model that is trained on the monolingual texts from a wide range of languages and has been shown to provide powerful sentence representations that lead to a new state-of-the-art performance on zero-resource cross-lingual language understanding tasks (Lample and Conneau, 2019; Pires et al., 2019).

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 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 heuristics have also been explored, such as matching tokens based on their surface forms (Feng et al., 2004; Samy et al., 2005; Ehrmann et al., 2011; Jain et al., 2019). A major problem in these methods is that the projections are produced independent of the sequence labels. By contrast, our method does not rely on heuristic projections, but models label projection through an attention model that can be jointly trained with other model components on the machine translated data.

3 Data

One of the most popular data sets for multilingual NLU with human translation is the ATIS data set (Price, 1990) and its multilingual extension (Upadhyay et al., 2018). The ATIS data set 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 cross-lingual transfer across a broader spectrum of language distances, we create the MultiATIS++ corpus[2] by extending both the training and test set of the English ATIS corpus to six additional languages. The resulting corpus covers nine languages in four different language families: Indo-European (English, Spanish, German, French, Portuguese, and Hindi), Sino-Tibetan (Chinese), Japonic (Japanese), and Altaic (Turkish).

For each of these languages, we hire professional native translators to translate the utterances and annotate the slots at the same time. When translating, the translators are required to preserve the meaning and structure of the original

---

[1] https://catalog.ldc.upenn.edu/LDC2019T04

[2] Corpus is under review by the LDC. Please contact saabm@amazon.com to obtain a copy.

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: MultiAtis++ parallel utterances with corresponding annotation example. The English utterance is human 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.

English sentences as much as possible. For example, repetitions such as "a flight before before 6 pm" are mimicked in the target language if possible. Slots spanning multiple tokens are marked using the BIO tagging scheme. We show an English training example and its translated versions in the other eight languages in Figure 1. Note that sub token slot values are tokenized from the surrounding text. For example in French, *d'Atlanta* is annotated with *d'{Atlanta|from_loc.city_name}* and the substring *d'* is not part of the slot value. Therefore, we tokenize *d'* and generate {*d'|O*} {*Atlanta|from_loc.city_name*}.

We report the data statistics in Table 1. 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 in front 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

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

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 $\boldsymbol{h}_{1...T}$

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

We explore two different encoder models:

- **biLSTM:** 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 in 104 languages. We take the hidden states of the top layer as the encoder representations and fine-tune the model on the NLU data.

4.2 Problems in Label Projection for Cross-Lingual NLU

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	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.

language. Prior work projects the slot labels using word alignments from statistical MT models or attention weights from neural MT models (Yarowsky et al., 2001; Schuster et al., 2019). The slot projection is done as a preprocessing step to prepare the training data for the downstream task.

Despite their empirical success, the label projection methods suffer from a major drawback: the projections are produced independently of the slot labels and the downstream task and are potentially erroneous. Jain et al. (2019) show that improving the quality of projection leads to significant improvements in the final 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. In addition, they incorrectly assume that each word in the target translation can be hard-aligned to a single word in the English sentence disregarding the morphological differences among languages.

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

4.3 Soft-Alignment via Attention

Given a source (English) utterance $s_{1...S}$ of length S and its translation $t_{1...T}$ 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 $e_{1...S}^{(src)}$.

Next, it encodes the translation $t_{0...T}$ (t_0 is inserted as the classification token) into a sequence of contextualized representations $h_{0...T}^{(tgt)} = \Theta_{enc}(t_{0...T})$, where Θ_{enc} is 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 $h_0^{(tgt)}$

$$p_{intent}(\cdot|t) = \text{softmax}(W_I h_0^{(tgt)} + b_I)$$

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

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

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

$$z_i = \sum_{j=1}^{T} a_{ij} h_j^{(tgt)}$$

where z_i is the hidden state at 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 $e_i^{(src)}$ and the key vector $h_j^{(tgt)}$ with learnable parameters to d dimensions. We then perform scaled dot-product attention on the projected query and key vectors

$$a_i = \text{softmax}\left(\frac{(e_i^{(src)} W^Q)(h^{(tgt)} W^K)^\top}{\sqrt{d}\tau}\right)$$
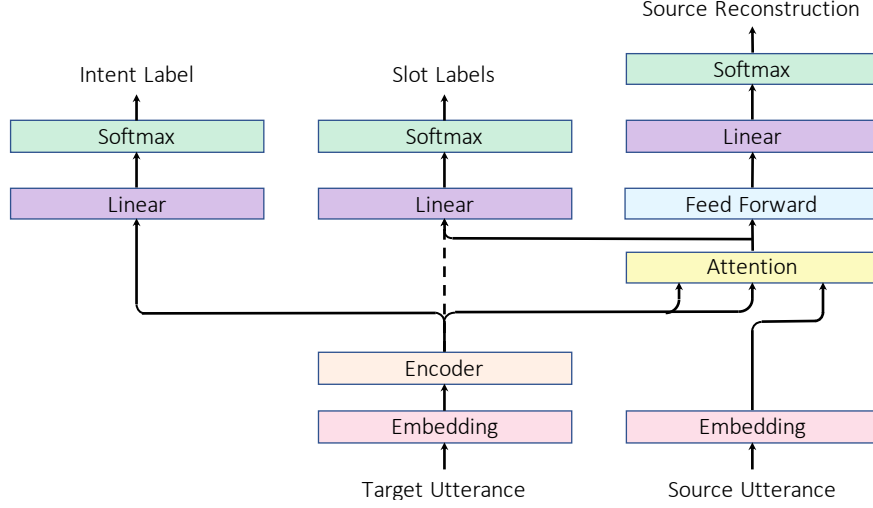
Figure 2: Architecture of our proposed model. The model is trained without external label projections: it learns to soft-align the representations of the target utterance to the source slot labels. The dotted line denotes the path during inference, where we directly connect the encoder module to the intent and slot classification layer to make predictions on the target utterance.

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 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 data with

objective $\mathcal{L}$ and the source language data with the supervised objective.

Note that 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 Cross-Lingual Transfer

In our first set of experiments, we explore using pre-trained multilingual BERT encoder with different training strategies to leverage the full training data for cross-lingual transfer:

- **Target only:** Using only the training data in the target language.

- **Cross-lingual:** Joint training on the concatenation of training data in all languages.

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. We set the initial learning rate to 1e-3 for the LSTM model and 1e-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

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
Cross-lingual	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
Cross-lingual	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 the full training data averaged over 5 runs. The *Target only* models are trained only on the target language training data. The *Cross-lingual* models are trained on the concatenation of training data in all languages.

model from (Devlin et al., 2019). Following (Goo et al., 2018) we use intent accuracy and slot F1 as evaluation metrics.

Results   Table 2 shows the results using the full supervised data averaged over 5 runs. Multilingual BERT encoder brings significant[3] 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. In addition, cross-lingual training with biLSTM on all the languages brings significant improvements on Hindi and Turkish over the target-only models – it improves intent accuracy by 2–5% and slot F1 by 4–11%. Further, BERT boosts the performance of cross-lingual training by 1–7% on intent accuracy and 1–9% on slot F1.

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 cross-lingual BERT model achieves higher F1 by +6.1% on Hindi and +7.1% on Turkish.

6   Zero-Shot Transfer Learning and Learning Curves

In this section, we compare the following methods for cross-lingual transfer where only a small

amount of annotated data or even no data (zero-shot) is available for the target language.

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

- **MT+project:** Using AWS Translate to translate the English data to the target language and word alignment from *fast-align*[4] for label projection; Joint training on the English and machine translated data.

- **MT+soft-align:** Using AWS Translate to translate the English data to the target language and our soft-alignment method described in Section 4.3; Joint 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 in the target language in this setting. We set the temperature τ of the attention module to 0.1.

6.1   Zero-Shot Results

As shown in Table 3, when training without MT data, BERT boosts the performance over LSTM by 14–32% on intent accuracy and 29–61% on slot F1 on all languages but Turkish – a dissimilar language to English. For both LSTM and BERT models, adding MT data with projected slot labels brings significant improvements on intent accuracy – by 13–33% for LSTM and 1–24% for BERT. However, we observe different trends on slot F1

---

[3] All mentions of statistical significance are based on a paired Student t test with $p < 0.05$.

[4] https://github.com/clab/fast_align

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 + project	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 + 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 + project	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 + 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 + project* rows correspond to models trained on the English and machine translated data with automatically projected slot labels using *fast-align*. The *MT + soft-align* row is the model trained on the English and machine translated data using our soft-alignment method.

for different languages. For example, when using BERT, adding MT data improves slot F1 by 20–44% on Chinese, Japanese, and Hindi, but hurts by around 6% on French. This may be due to its mixed effect – training directly on the target language data is beneficial especially for languages dissimilar to English, however, the noisy projection of the slot labels could also bring harm to the model.

Next, we compare our soft-alignment method with others. First, our method is more robust across languages than the projection method – it achieves consistent improvements over the BERT models trained without translated data in terms of both metrics, while the projection method leads to a degradation on French and Turkish. Next, we compare our method with the projection method in terms of slot F1. Although we observe a small degradation on Spanish and Portuguese, our method achieves significant improvements on five of the eight languages, especially on French (+9.5%), Hindi (+9.1%), and Turkish (+38.1%).

To further analyze the slot F1 performance of MT+project in comparison to MT+soft-align, we measure the slot projection accuracy. The projection accuracy is calculated by the matching rate between token-slot pairs in the human translated data versus the MT+project data. We find that the projection accuracy is 20% and 57% for Turkish and Hindi correspondingly while for the other languages the accuracy is around 70%. The low projection accuracy for Turkish can be attributed to the rich morphology of the language. For example for

Intent acc.	Average over 8 languages
MT + soft-align	94.87
- Reconstruction loss	94.95
- Source	94.64
Slot F1	Average over 8 languages
MT + soft-align	80.00
- Reconstruction loss	76.42
- Source	71.62

Table 4: Ablation results for zero-shot transfer learning averaged over 8 languages. Languages include es, de, zh, ja, pt, fr, hi, tr.

the Turkish phrase *san francisco'ya*, in MT+project *san* is not tagged as a city_name and the whole token *francisco'ya* is tagged with city_name rather than the substring *francisco*. Even though both MT+project and MT+soft-align use WordPiece internally, MT+soft-align benefits from the consistent end-to-end training to handle rich morphology.

6.2 Ablation Study

Table 4 shows the impact of dropping different components of the soft-alignment architecture. For intent accuracies, the components have no impact as the intent class is on the utterance level and alignment has no additional benefit. On the other hand, the soft-alignment components are crucial for slot labeling. Dropping the reconstruction loss results in a 3.6% drop, while dropping the source utterance is more severe and results in 8.4% degradation.
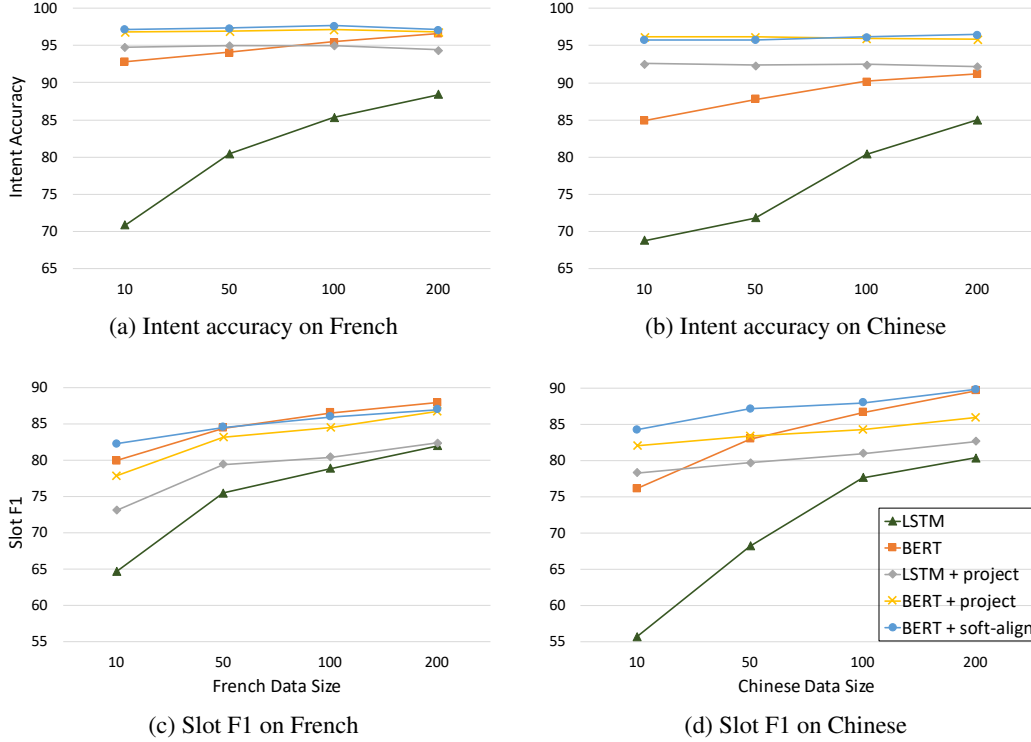
Figure 3: Results for different sizes of the target language training data averaged over 5 runs. *LSTM* and *BERT* correspond to the *No MT* setup while the other models use MT data.

6.3 Learning Curves

Results for cross-lingual transfer using different sizes of the target language training data are shown in Figure 3. We select French as a similar language to English, and Chinese as a dissimilar language. We find that: 1) BERT achieves remarkably good performance in all data scales when trained both with and without machine translated data. 2) BERT brings promising results given several hundred training examples in the target language: BERT trained without MT achieves comparable slot F1 score to the best translation method on Chinese and achieves higher score on French. 3) Our soft-alignment method is more robust than the projection method given a small amount of target training data. For example, on Chinese, the slot F1 score of the projection method quickly plateaus as the target data size increases, while our method leads to continuous improvements.

7 Conclusion

We introduce MultiATIS++, a multilingual NLU corpus that extends the Multilingual ATIS corpus to nine languages across four language families. We use the corpus to explore three different cross-lingual transfer methods: a) using the multilingual

BERT encoder, b) the traditional projection method based on machine translation and slot label projection using external word alignment tools, and c) our proposed soft-alignment method that requires no label projection as it performs soft label alignment dynamically during training via an attention mechanism. While the traditional projection method obtains comparable results on intent detection, it relies heavily on the quality of the slot label projection. Experimental results show that our method improves slot F1 over the traditional method on five of the eight languages, especially on Turkish, which is known to be different from other languages in terms of morphology.

In addition, we find that using multilingual BERT brings substantial improvements for both the cross-lingual and zero-shot setups, and that given a small amount of annotated data in the target language, multilingual BERT achieves comparable or even higher scores than the translation methods on slot F1.

We hope that our work will encourage more research on learning cross-lingual sentence representations to match the performance of translation approaches as well as more advanced cross-lingual transfer methods that are robust to the translation and label projection errors.

References

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)*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

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

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? *CoRR*, abs/1906.01502.

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.

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.