Natural Language Processing (2024), 1–16
doi:10.1017/nlp.2024.17

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Intent detection and slot filling for Persian: Cross-lingual training for low-resource languages

Reza Zadkamali, Saeedeh Momtazi and Hossein Zeinali

Amirkabir University of Technology, Tehran, Iran.
Corresponding author: Saeedeh Momtazi; Email: momtazi@aut.ac.ir

(Received 30 November 2022; revised 28 May 2023; accepted 20 August 2023)

Special Issue on '**Natural Language Processing Applications for Low-Resource Languages**'

Abstract

Intent detection and slot filling are two necessary tasks for natural language understanding. Deep neural models have already shown great ability facing sequence labeling and sentence classification tasks, but they require a large amount of training data to achieve accurate results. However, in many low-resource languages, creating accurate training data is problematic. Consequently, in most of the language processing tasks, low-resource languages have significantly lower accuracy than rich-resource languages. Hence, training models in low-resource languages with data from a richer-resource language can be advantageous. To solve this problem, in this paper, we used pretrained language models, namely multilingual BERT (mBERT) and XLM-RoBERTa, in different cross-lingual and monolingual scenarios. To evaluate our proposed model, we translated a small part of the Airline Travel Information System (ATIS) dataset into Persian. Furthermore, we repeated the experiments on the MASSIVE dataset to increase our results' reliability. Experimental results on both datasets show that the cross-lingual scenarios significantly outperform monolinguals ones.

Keywords: intent detection; slot filling; Persian language understanding; joint learning; low-resource languages

1. Introduction

Dialog systems are widely used in various personal assistants such as Google Assistant, Amazon Alexa, Apple Siri, and Microsoft Cortana. Natural language understanding (NLU) plays an essential role in enabling users to accomplish their tasks through verbal interactions. NLU typically involves two tasks: intent detection and slot filling. In particular, intent detection aims to identify a speaker's intent from a given utterance which can be treated as a sentence classification problem. In contrast, slot filling extracts the correct argument values from the utterance for intents slots, which can be treated as a sequence labeling task that maps an input word sequence into the corresponding slot tags sequence. Several joint learning methods have been proposed to improve performance over independent models to model and exploit the relationship between intent detection and slot filling. Figure 1 demonstrates a typical sample from the Airline Travel Information System (ATIS) (Tur, Hakkani-Tür, and Heck 2010) training set in which the slot filling task is labeled with the IOB representation.

NLU is eventually required in many languages, most of which do not have large annotated training datasets. In response to the lack of human-labeled data, various methods were developed to train general-purpose language representation models from a large set of unannotated texts, such as Word2Vec (Mikolov *et al.* 2013) and Glove (Pennington, Socher, and Manning 2014).

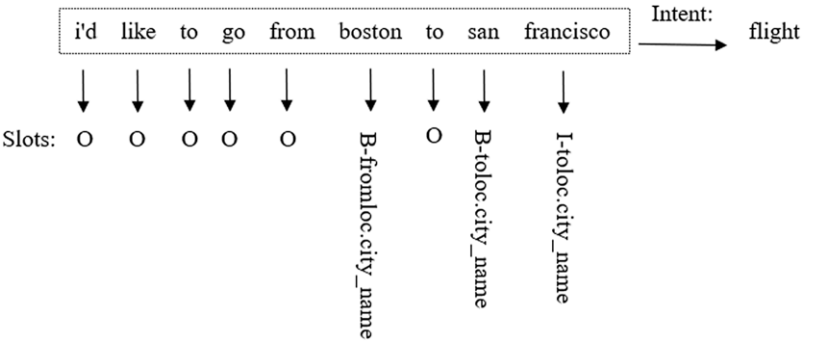

Figure 1. An example of intent and slot labels from the ATIS dataset.

Pretrained models can be fine-tuned on natural language processing (NLP) tasks and have significantly improved over training on task-specific annotated data. More recently, contextualized word representations such as ELMo (Peters *et al.* 2018), Generative Pre-trained Transformer (GPT) (Radford *et al.* 2018), and Bidirectional Encoder Representations from Transformers (BERT) (Kenton and Toutanova 2019) were proposed and have created state-of-the-art models for a wide variety of NLP tasks. Although these pretrained models significantly improved the performance of different NLP tasks, fine-tuning the models with labeled data is still an important issue.

An aspect of generalizability refers to whether a model can be applied outside of the language in which it has been trained. Therefore, a transfer learning approach from a rich-resource language to a low-resource language would be desirable. In addition, developing cross-lingual transfer methods for intent detection and slot filling is challenging due to the lack of multilingual datasets that have been annotated according to the same guidelines. In this work, we use the ATIS dataset, which contains thousands of English utterances (the rich-resource data), and a novel dataset of Persian utterances (the low-resource data), annotated according to a similar annotation scheme. These data make it possible to examine cross-linguistic transfer learning methods from rich-resource language to a low-resource language. We aim to investigate if it is feasible to achieve a cross-lingual joint model using multiple phases of fine-tuning in training while outperforming the monolingual models. To summarize, the key contributions of this paper are as follows:

- To produce our low-resource data, we manually translated and annotated 1462 samples of the ATIS dataset from English to Persian.
- We explore the performance of current cross-lingual pretrained language models such as multilingual BERT (mBERT) (Kenton and Toutanova 2019) and XLM-RoBERTa (Conneau *et al.* 2020) to address the lack of multilingual human-labeled data.
- We fine-tuned the JointBERT+CRF (Chen, Zhuo, and Wang 2019) model with different scenarios in the training phase and reported the results.

The structure of the paper is as follows: in Section 2, we review and discuss related research works concisely. In Section 3, our proposed model and training scenarios are presented. The dataset statistics, evaluation metrics, and experiments setup are explained in Section 4. Detailed experimental results and analysis are given in Section 5, and finally, our conclusions are presented in Section 6.

2. Related works

2.1 Related works on joint intent detection and slot filling models

Pipelining the two subtasks has produced accurate results but is prone to error propagation. Two tasks, slot filling and intent detection, can be solved simultaneously by a joint model.

https://doi.org/10.1017/nlp.2024.17 Published online by Cambridge University Press

Table 1. Summary of slot filling and intent detection results of existing models on SNIPS and ATIS datasets

	SNIPS			ATIS		
Models	F1	Accuracy	Exact match	F1	Accuracy	Exact match
Zhang et al. (2019)	91.8	97.3	80.9	95.2	95.0	83.4
Qin et al. (2019)	94.2	98.0	86.9	95.9	96.9	86.5
Zhu et al. (2020)	96.44	99.14	—	95.79	98.43	—
Yang et al. (2021)	95.2	98.7	88.9	96.0	97.0	87.1
Wang et al. (2018)	—	—	—	96.89	98.99	—
Tang et al. (2020)	97.2	**99.7**	**93.6**	96.4	**99.0**	**89.6**
Chen et al. (2019)	**98.6**	97.0	92.8	**97.5**	96.1	88.6

In many cases, joint distributions implicitly model via joint loss back-propagation (Guo *et al.* 2014; Hakkani-Tür *et al.* 2016; Liu and Lane, 2016). It would be beneficial to capture the relationship between intents and slots explicitly (Han *et al.* 2021).

A joint capsule neural network is proposed by Zhang *et al.* (2019), which uses a dynamic routing-by-agreement scheme between capsule layers. In dynamic routing-by-agreement, it explicitly models words, slots, and intents at the utterance level.

Qin *et al.* (2019) performed the token-level intent detection to improve the robustness of intent detection and proposed a stack-propagation framework that incorporates intent information to guide the slot filling.

Zhu *et al.* (2020) used a dual model for joint intent detection and slot filling to generate sentences based on structured semantic forms. They performed a novel framework for semi-supervised NLU by incorporating the dual model in order to take advantage of unlabeled data.

Yang *et al.* (2021) proposed a joint model of intent detection and slot filling based on a position-aware multi-head masked attention mechanism. The explicit feature interactions are modeled as the inner product of the word encoding vector and the intent-slot feature vectors. Wang *et al.* (2018) applied a new bidirectional recurrent neural networks (Bi-RNN) model to jointly perform the intent detection and slot filling tasks by considering their cross-impact using two correlated bidirectional long short-term memories (Bi-LSTM). Tang *et al.* (2020) proposed a graph-based conditional random field (CRF) for modeling the implicit connections within slot–slot and slot–intent pairs and solved the incompatibility between slot tags and intent tags by employing a mask mechanism. Chen *et al.* (2019) proposed to use the contextual BERT model to learn the two tasks jointly.

Table 1 presents an overview of the results of the existing models and reports F1 score, accuracy, and exact match (EM) on both SNIPS and ATIS datasets.

2.2 Related works on cross-lingual models

English can be used in conjunction with low-resource languages to address the lack of annotated data, which has become a popular topic recently. Castellucci *et al.* (2019) considered transfer learning from English to Italian. Upadhyay *et al.* (2018) leveraged multilingual word embeddings that share a common vector space across various languages to do zero-shot and almost zero-shot transfer learning in intent detection and slot filling. They translated the ATIS English dataset into Turkish and Hindi. Xu *et al.* (2020) proposed an end-to-end approach for jointly aligning and predicting target slot labels for cross-lingual transfer. They released Multi-ATIS++, a

https://doi.org/10.1017/nlp.2024.17 Published online by Cambridge University Press

multilingual NLU corpus with six new languages: Spanish, German, French, Portuguese, Chinese, and Japanese.

Artetxe *et al.* (2017) developed a method to decrease reliance on large bilingual dictionaries using smaller seed dictionaries. Their approach involved a self-learning framework that can be used in conjunction with any dictionary-based mapping technique. They learned how to map source and target word embeddings through a small word dictionary. He *et al.* (2020) examined the effectiveness of dividing slot tagging models into the language-shared part and language-specific parts to transfer cross-lingual knowledge and improve monolingual slot tagging. Moreover, they refined shared knowledge with language discriminators and reinforce information separation through adversarial training.

Gritta and Iacobacci (2021) used translated task data to encourage the model to generate similar sentence embeddings for different languages. Gritta *et al.* (2022) introduced CrossAligner, a cross-lingual transfer method that converts training data in English into a task that can be applied to any language. This task is used to synchronize model predictions across different languages. They have also presented a contrastive alignment method that reduces the cosine distance between translated sentences while increasing it for unrelated sentences. This new method requires significantly less data compared to previous works. Additionally, they have suggested Translate-Intent as a simple and efficient baseline approach that surpasses previous Translate-Train methods without using error-prone data transformations like slot label projection.

Schuster *et al.* (2019) studied transfer to low-resource languages, from English to Spanish and Thai. Three methods of cross-linguistic transfer have been used: translation of training data, cross-lingual pretrained embeddings, and a multilingual machine translation encoder as contextual vectors (CoVe) (McCann *et al.* 2017) for word representations. Liu *et al.* (2019) proposed an attention-informed mixed-language training instead of manually selecting the word pairs, and they proposed to extract source words based on the scores computed by the attention layer of a trained English task-related model and then generate word pairs using existing bilingual dictionaries. Qin *et al.* (2021) proposed an augmentation framework to generate multilingual code-switching data to fine-tune mBERT for aligning representations from rich-resource and multiple low-resource languages by mixing their context information.

Using cross-lingual transfer, the model outperforms training on limited data from the low-resource language. However, some challenges remain to deal with, for example, aligning intents and slots from source and target languages in different scenarios, such as differences in syntax or grammar between languages or translation gaps where certain words may not have an equivalent translation. Additionally, idiomatic expressions, cultural differences, and regional dialects can make it challenging to align intents and slots, as they can vary significantly between languages and regions. Moreover, ensuring model generalizability across different languages and language families remains an area of research that needs exploration.

These challenges motivated us to propose model for cross-lingual intent detection and slot filling which benefit from the advantages of large amount of data in rich-resource languages while still using limited data from the target low-resource language to better learn the features of the target language.

3. Proposed model

3.1 JointBERT+CRF

As mentioned, we used the JointBERT+CRF model to test different scenarios. The overall structure of the model is presented in Figure 2. In this figure, $FFNN_{ID}$ and $FFNN_{SF}$ denote the feed forward neural networks which consists of a single linear layer and are used in the last layer of the architecture for intent detection and slot filling, respectively. As shown in Figure 2, it consists of three layers: an encoding layer and two decoding layers of intent detection and slot filling.

https://doi.org/10.1017/nlp.2024.17 Published online by Cambridge University Press

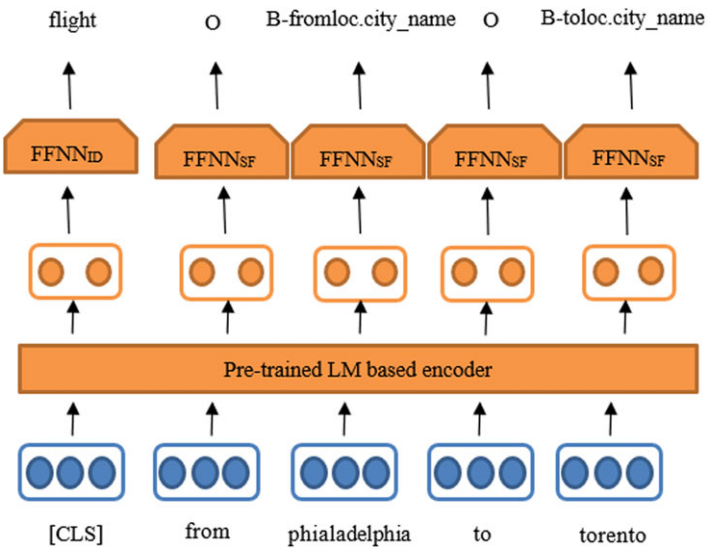

Figure 2. Illustration of the JointBERT+CRF model. The input query is "from Philadelphia to Toronto" (Chen *et al*. 2019).

Encoding layer: A pretrained multilayer bidirectional transformer encoder named BERT is employed in the encoding layer to produce contextualized latent feature embeddings. Tokens are inserted as follows: a classification embedding ([CLS]) as the first token and a unique token ([SEP]) as the last token. For a sequence of input tokens $x = (x^1,\ldots,x^T)$, the output of BERT is $H= (h_1,\ldots,h_T)$. In the BERT model, two strategies are employed to pretrain the model on large-scale unlabeled texts: the masked language model (MLM) and next sentence prediction (NSP). The cross-lingual BERT models, specifically, mBERT (Kenton and Toutanova 2019) and XLM-RoBERTa (Conneau *et al*. 2020), provide powerful context-dependent sentence representations that can be used for cross-lingual tasks, including intent detection and slot filling by fine-tuning.
Intent detection layer: The BERT model can easily be extended to a joint intent detection and slot filling model. We feed the hidden states of the unique token [CLS] denoted as h_1, into a feed forward layer and then passed to a softmax layer, The intent is predicted as follows:

$$y^i = softmax(W^i h_1 + b^i) \tag{1}$$

Slot filling layer: For slot filling, first we feed the final hidden states of other tokens $h_2,\ldots,h_T$ into a feed forward layer and then it will be passed into a softmax layer to classify over the slot labels. Since BERT tokenizes each input token into multiple sub-tokens by using WordPiece tokenization, we only use the hidden states corresponding to the first sub-token as the input to the slot decoder:

$$y_n^s = softmax(W^s h_n + b^s), n \in 1,\ldots,N \tag{2}$$

where h_n is the hidden state corresponding to the first sub-token of word x_n.
CRF layer: The predictions for slot labels are influenced by the surrounding words. As shown by Chen *et al*. (2019), structured prediction models, such as CRFs, can improve slot filling performance. In this case, we added CRF to model slot label dependencies on top of the joint BERT model. Given an input sentence x of length L and the tag scores y, the final score of a sequence of tags z is calculated as follows:

$$S(x,y,z) = \sum_{t=1}^{L} \left(A_{z_{t-1},z_t} + y_{t,z_t}\right) \tag{3}$$

https://doi.org/10.1017/nlp.2024.17 Published online by Cambridge University Press

In the transition matrix A, $A_{p,q}$ represents the binary score of transitioning from tag p to tag q, and y_{t,z_t} represents the unary score of assigning tag z to the t^{th} word. During the training phase, we aim to maximize the following objective function given the ground truth sequence of tags z:

$$\begin{aligned} O &= logP(z|x) \\ &= S(x, y, z) - log \sum_{\tilde{z} \in Z} e^{S(x,y,\tilde{z})} \end{aligned} \tag{4}$$

All paths for tagging can be represented by Z.
Joint training: The joint model has a training objective loss (L), which is the weighted sum of the intent detection loss L_{ID} and the slot filling loss L_{SF}:

$$\mathscr{L} = \lambda \mathscr{L}_{ID} + (1 - \lambda)\mathscr{L}_{SF} \tag{5}$$

The hyperparameter λ represents the combination weight: $0 < \lambda < 1$.

3.2 Training scenarios

Considering the cross-lingual approach in our proposed model, we aim to perform various training scenarios in order to train our models in English (EN) and Persian (PR). In all experiments, we consider that a large amount of samples are available in the English training data, while the Persian training data includes a small amount of samples.

The training scenarios are as follows:

- **PR:** In this model, we only use Persian training data, that is, the joint intent-slot model is fine-tuned in one step using the Persian data.
- **EN:** This model only uses English training data as our high-resource data.
- **PR→EN:** The model is first trained on the Persian training data and then on the English training data.
- **EN→PR:** The model is trained on English training data and Persian training data, respectively. Figure 3 provides a comprehensive view of this scenario.
- **EN + PR:** A combination of English and Persian training data is used to train the model. Figure 4 represents this scenario.

4. Evaluation

4.1 Dataset

To evaluate different scenarios, we used two benchmark datasets:

- **ATIS:** The ATIS is a popular and widely used dataset in NLU research, which contains English audio recordings of people making flight reservations. This dataset comprises 4478 utterances for training and 893 utterances for the test, containing 21 intent and 120 slot tags. In the training stage, we used English training ATIS utterances as our rich-resource dataset. To achieve the low-resource dataset, we translated 500 random utterances (approximately 10% of the original data) of ATIS from English to Persian; in the test stage, we translated the entire test set of the ATIS; in addition, we added 69 informal translated utterances.[a] Some examples of translated ATIS utterances with corresponding labels are shown in Figure 5.

[a]The Persian ATIS dataset is available at https://github.com/MobinZadkamali/Intent-Detection-and-Slot-Filling-for-Persian-Crosslingual-Training-for-Low-resource-Languages

https://doi.org/10.1017/nlp.2024.17 Published online by Cambridge University Press

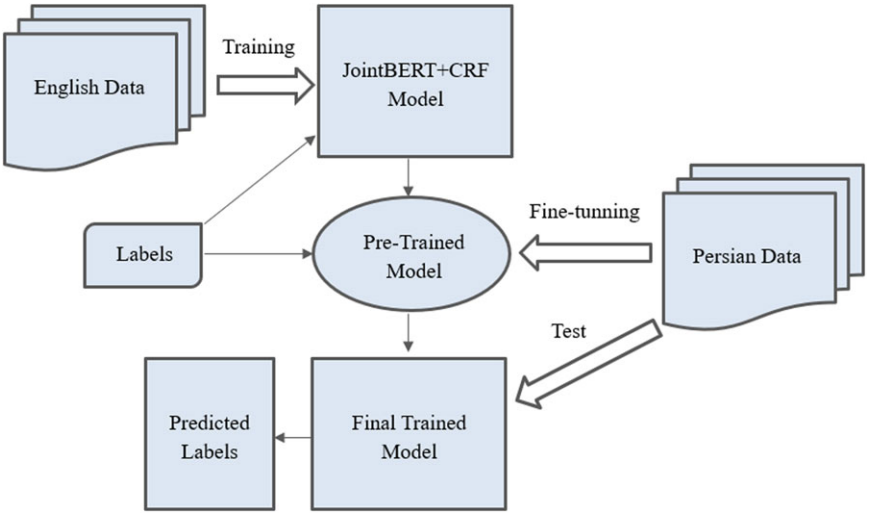

Figure 3. The architecture of the EN→PR training scenario; the model was first trained on English training data and then on Persian training data.

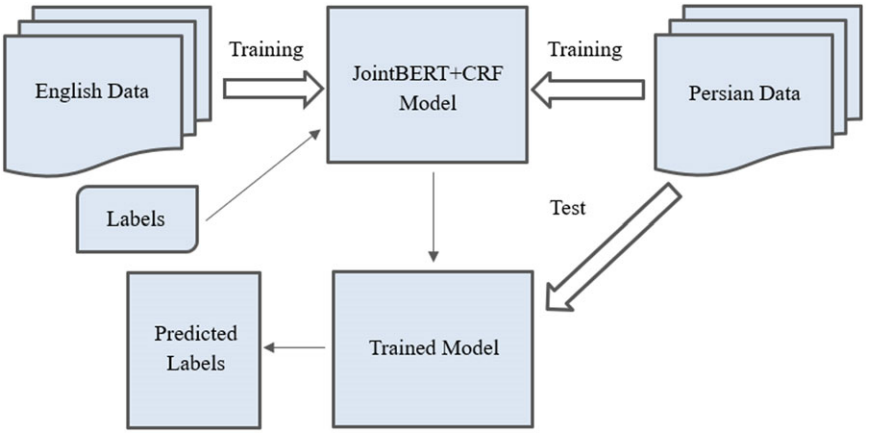

Figure 4. The architecture of the EN + PR training scenario; the model was trained on a combination of English and Persian training data.

- **MASSIVE:** MASSIVE is a newly released joint NLU dataset (FitzGerald *et al.* 2022) composed of one million realistic, parallel, labeled virtual assistant utterances spanning 51 languages, 18 domains, 60 intents, and 55 slots (108 IOB slot tags). MASSIVE contains 12664, 2974, and 2974 samples for training, development, and testing set, respectively. Similar to ATIS, we utilized the English MASSIVE training set and 10% of the Persian MASSIVE data for training, along with all of the Persian MASSIVE test sets for testing.

Statistics of both datasets are presented in Table 2.

4.2 Evaluation metrics

To evaluate different scenarios, we employed three standard evaluation metrics, including F1 score for slot filling, accuracy for intent detection, and EM for both intent detection and slot

https://doi.org/10.1017/nlp.2024.17 Published online by Cambridge University Press

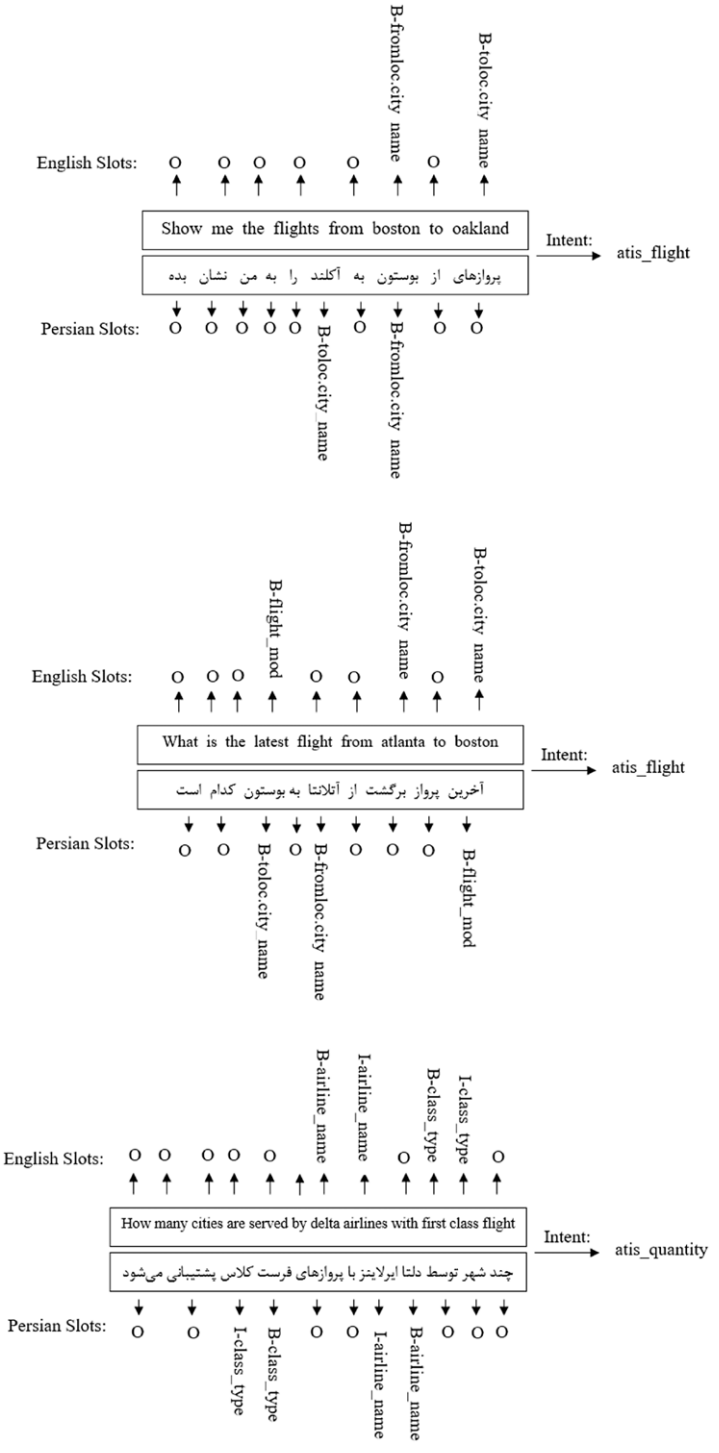

Figure 5. Examples of translated ATIS utterances with corresponding labels. The top part is the original English utterance, and the bottom part is the Persian translation.

https://doi.org/10.1017/nlp.2024.17 Published online by Cambridge University Press

Table 2. Datasets statistics

Dataset	Language	Vocab size	#Train	#Valid	#Test	#Slot	#Intent
ATIS	Persian	1428	500	481	481	130	26
	English	5473	4478	500	893	130	26
MASSIVE	Persian	16432	11514	2033	2974	108	60
	English	16432	11514	2033	2974	108	60

filling. EM is introduced to count the testing samples with absolutely correct prediction. The other metrics are computed by the equations below:

$$P_{slots} = \frac{S}{M} \tag{6}$$

$$R_{slots} = \frac{S}{N} \tag{7}$$

$$F1_{slots} = \frac{2 * P_{slots} * R_{slots}}{P_{slots} + R_{slots}} \tag{8}$$

$$Acc_{intents} = \frac{T}{K} \tag{9}$$

where N is the number of gold slot chunks in the test set, M is the number of predicted slot chunks, S is the number of correctly predicted slot chunks, T is the number of correctly predicted intents, and K is the number of utterances.

4.3 Experiments setup

We conduct experiments on our datasets to study the usefulness of pretrained language model-based encoders. Here, we employ XLM-RoBERTaBASE (Conneau *et al.* 2020) and mBERT (Kenton and Toutanova 2019) (two recent state-of-the-art pretrained language models that support Persian) as the encoders.

- **mBERT:** mBERT is a BERT multilingual model with 12 layers, 768 hidden units each, 12 attention heads, pretrained on the top 104 languages (including Persian) with texts from Wikipedia using MLM and NSP objectives. The entire model has 110 M parameters.
- **XLM-RoBERTa:** XLM-RoBERTa is a multilingual variant of RoBERTa, pretrained on a 2.5TB multilingual dataset on 100 languages (including Persian). It does not use the NSP task for training and is only trained using the multilingual MLM.

The maximum length of an utterance is 50. The batch size is set as 128. Adam is used for optimization with an initial learning rate of 5e-5. The dropout probability is 0.1. In the case of training over one language or the mixture of two languages, the maximum number of epochs is 20. If training is done in two phases over two languages (each phase for one language), the maximum number of epochs is 20 for each phase. The reported results are the average of five runs using five different random seeds.

https://doi.org/10.1017/nlp.2024.17 Published online by Cambridge University Press

Table 3. Experimental results with all the scenarios, using mBERT pretrained language model as the encoder on ATIS and MASSIVE test dataset

	ATIS			MASSIVE		
Strategy	F1	Accuracy	Exact match	F1	Accuracy	Exact match
EN	50.96	86.48	17.25	79.68	87.35	69.43
PR	63.54	76.29	28.89	53.15	32.14	16.57
PR→EN	73.59	**90.64**	47.19	**79.88**	**87.79**	**69.87**
EN→PR	74.58	90.22	48.44	79.01	86.68	68.72
EN + PR	**75.59**	90.22	**50.1**	79.61	87.62	69.36

Table 4. Experimental results with all the scenarios, using XLM-RoBERTa pretrained language model as the encoder on ATIS and MASSIVE test dataset

	ATIS			MASSIVE		
Strategy	F1	Accuracy	Exact match	F1	Accuracy	Exact match
EN	3.18	46.77	0.0	67.26	**84.76**	58.8
PR	39.78	70.27	8.1	43.23	50.2	25.15
PR→EN	16.56	74.63	1.24	63.99	83.01	55.64
EN→PR	57.77	**79.0**	**23.07**	65.05	82.64	56.82
EN + PR	**58.56**	78.17	21.82	**68.35**	83.82	**59.91**

5. Results

Table 3 presents the results of all training scenarios using mBERT as the encoder. The table shows that cross-lingual scenarios have been more effective, achieving better results than monolingual scenarios.

In the scenario that we only use English data, all evaluation metrics in the MASSIVE dataset and the accuracy metric for intent detection in the ATIS dataset outperforms the scenario in which only the Persian dataset has been used. The reason for this could be the larger size of English training datasets. Among all scenarios, the PR→EN exhibits the best results in the MASSIVE dataset. In the ATIS dataset, the best results have been obtained in two scenarios, PR→EN and EN + PR, which indicates that the combination of the rich-resource English dataset and the low-resource Persian dataset has been effective.

In Table 4, we have been given results for all training scenarios using XLM-RoBERTa as the encoder. As shown in this table, on the ATIS dataset, in the case that our training data are only Persian data, better performance has been achieved in all evaluation metrics than in the case where our training data are English data. However, on the MASSIVE dataset, similar to mBERT encoding results, the large size of the MASSIVE English training dataset led to better performance.

According to our experiments, the mBERT pretrained language model yields the best results for both datasets. The better performance of mBERT in contrast to XLM-RoBERTa language model can be due to the curse of multilinguality (Conneau *et al.* 2020). In the ATIS dataset, the highest value of F1 was obtained in the EN→PR scenario (75.94), the highest accuracy value was achieved in the PR→EN scenario (90.64), and the highest EM value was obtained in the EN + PR mode (50.1). In the MASSIVE dataset, the highest value of all three metrics was attained in the PR→EN

https://doi.org/10.1017/nlp.2024.17 Published online by Cambridge University Press

Table 5. Experimental results on the MASSIVE dataset using an equal number of Persian and English samples for training on the mBERT and XLM-RoBERTa language models

	mBERT			XLM-RoBERTa		
Strategy	F1	Accuracy	Exact match	F1	Accuracy	Exact match
PR	79.31	**87.79**	69.36	68.91	85.44	59.27
PR→EN	**80.6**	**87.79**	**71.47**	**71.13**	**86.17**	**61.97**
EN→PR	80.22	87.65	70.0	71.06	85.93	61.43
EN + PR	80.12	87.62	70.24	71.06	85.93	**61.97**

scenario (the obtained values for F1, Accuracy, and EM metrics, respectively, are equal to 79.88, 87.79, and 69.87).

5.1 Equal multilingual resources

In order to compare the performance of 10% of the Persian dataset with the comparable data in two languages, we used all 12664 Persian samples of the MASSIVE dataset. The results of the mBERT and XLM-RoBERTa language models have been presented in Table 5.

Not surprisingly, the performance of the MASSIVE dataset's results in the PR scenario on mBERT language model improved by 25.68% on slots F1, 35.24% on intents accuracy, and 34.12% on EM, and when mBERT has been replaced by XLM-RoBERTa, we have been witnessing 26.16%, 55.65%, and 52.79% improvement on slots F1, intents accuracy, and EM, respectively. However, in the case of using mBERT as our pretrained language model, our experiments only obtained an improvement of 0.72% in slots F1 and 1.6% in EM and no improvement in intents accuracy for cross-lingual scenarios. In the case of using XLM-RoBERTa, the improvements in comparison to the best results of using 10% of the Persian dataset with XLM-RoBERTa are as follows: 2.78% on slots F1, 1.41% on intents accuracy, and 2.06% EM. In Figure 6, we investigate the performance variation when utilizing different percentages of Persian data in the PR→EN scenario. As it can be seen from the figure, increasing the size of the Persian dataset does not improve EM metric significantly.

5.2 Error analysis

As shown in Figure 7 and Figure 8, we demonstrate two typical testing examples, one from the ATIS dataset annotated by EN+PR training model and another one from the MASSIVE dataset annotated by PR→EN training model. Both of the samples obtained the right slot filling and intent detection results by using the mBERT language model.

As shown in Figures 9 and 10, we gather two examples of each dataset for which our model fails. In the sample of Figure 9, our model obtains the wrong slot tag 'depart_day.day_time' for the word 'tenth' instead of 'daepart_day.dayname'. In the sample of Figure 10, our model annotates the words 'US dollar' and 'Iranian rial' as slot 'currency_name'. However, they actually represent the slot 'news_topic'. Additionally, our model incorrectly assigns the intent tag 'qa_currency' instead of 'news_query'. Nevertheless, even human beings may find it challenging to recognize the correct slots and intent.

Table 6 presents the confusion matrix for the intent detection task of our ATIS dataset. There are a few errors due to the imbalanced data problem since most utterances are labeled as 'flight'. The intent labels such as 'airline;flight_no' and 'flight;airline' have significantly less representation

https://doi.org/10.1017/nlp.2024.17 Published online by Cambridge University Press

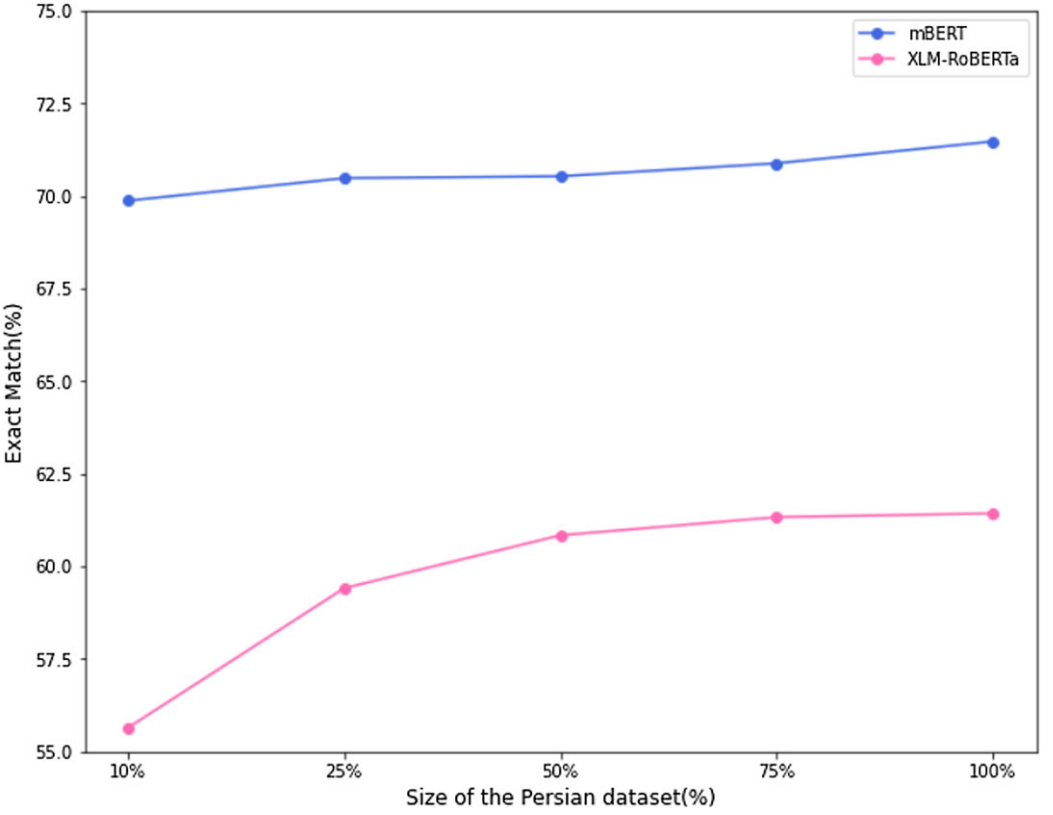

Figure 6. Performance over different size of Persian data for training phase in PR→EN scenario.

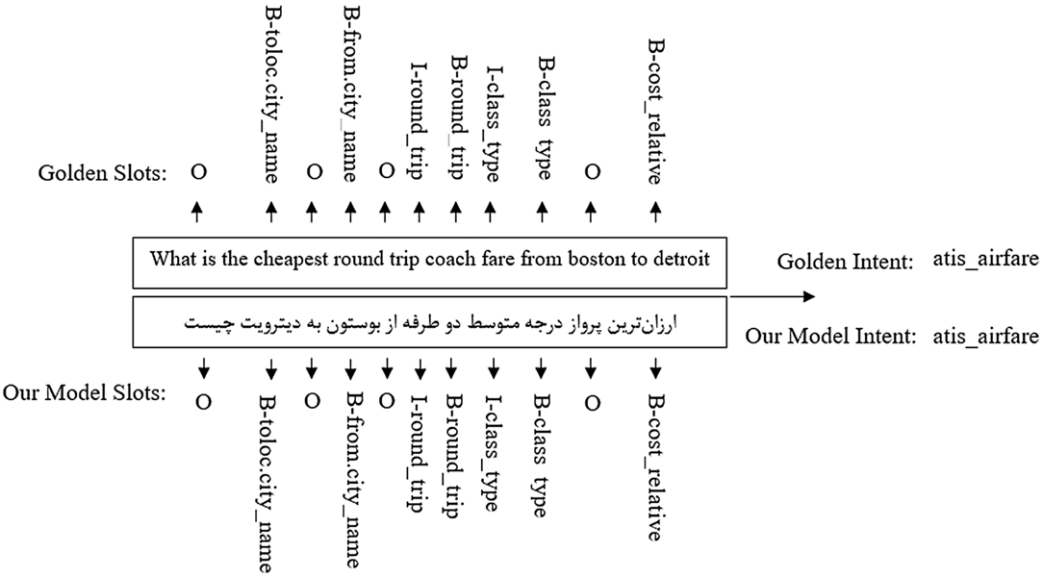

Figure 7. Demonstration of a test sample from the ATIS dataset generated by the PR + EN scenario and mBERT model. The gold label and translation are also included.

https://doi.org/10.1017/nlp.2024.17 Published online by Cambridge University Press

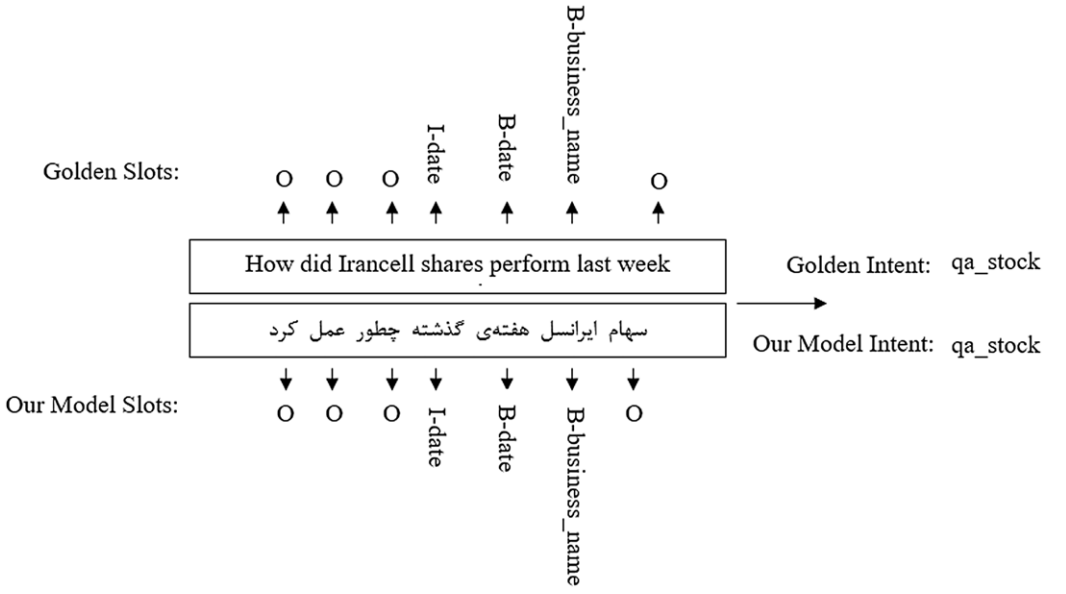

Figure 8. Demonstration of a test sample from the MASSIVE dataset generated by the PR→EN scenario and mBERT model. The gold label and translation are also included.

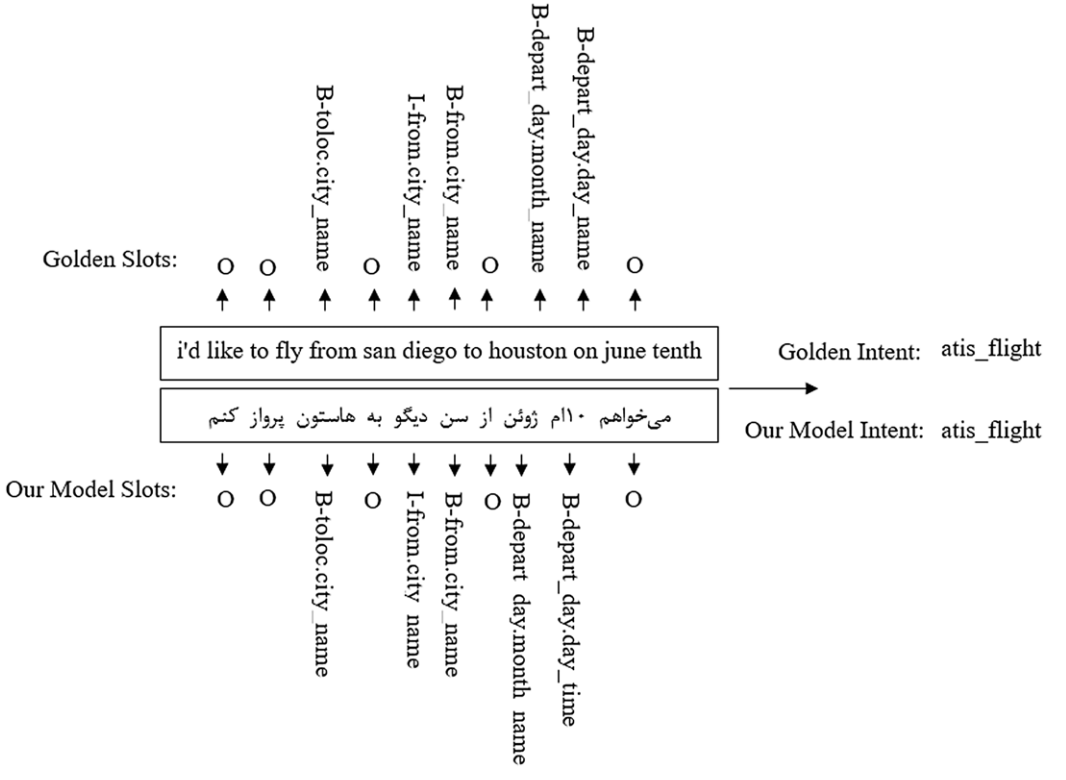

Figure 9. Demonstration of a test sample from the ATIS dataset showing incorrect output of our model with the EN + PR scenario. The golden label and translation are also included.

https://doi.org/10.1017/nlp.2024.17 Published online by Cambridge University Press

Table 6. Confusion matrix for intent detection of ATIS dataset

Correct-estimated	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q
a.flight	322	0	4	6	1	0	2	0	2	0	0	0	0	0	0	0	0
b.flight_time	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c.airfare	2	0	28	1	0	0	0	0	0	0	0	0	0	0	0	0	0
d.aircraft	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
e.ground_service	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0
f.airport	0	0	0	0	0	10	0	0	0	0	0	3	0	0	0	0	0
g.airline	1	0	2	0	1	0	17	0	0	0	0	0	0	0	0	0	0
h.distance	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
i.abbreviation	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0
j.ground_fare	0	0	2	0	1	0	0	0	0	1	0	0	0	0	0	0	0
k.quantity	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
l.city	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
m.flight_no	3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
n.capacity	0	0	0	0	1	0	0	0	0	1	0	0	0	11	0	0	0
o.meal	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
p.flight;airfare	1	0	5	0	0	0	0	0	0	0	0	0	0	0	0	2	0
q.airline;flight_no	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

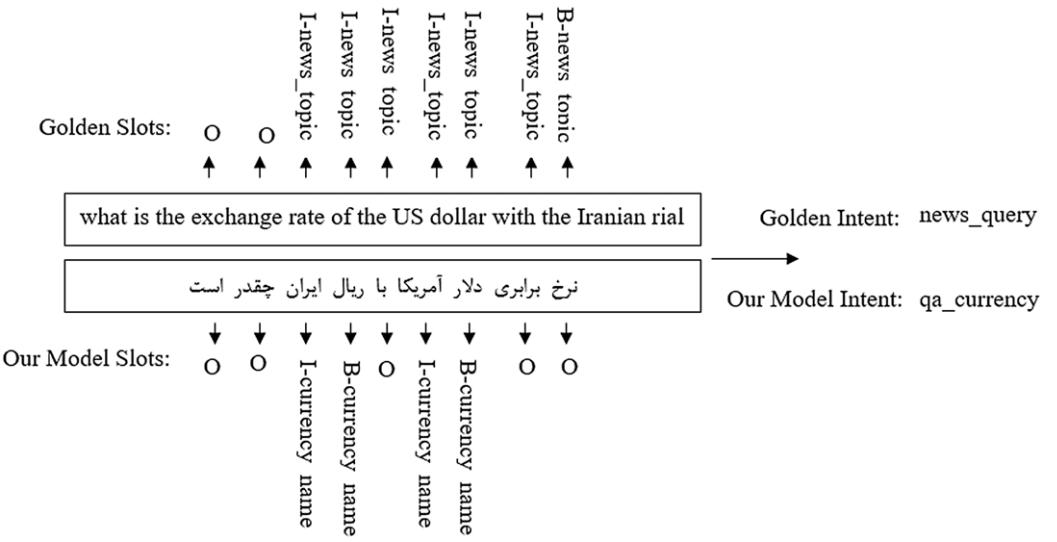

Figure 10. Demonstration of a test sample from the MASSIVE dataset showing incorrect output of our model with the EN→PR scenario. The golden label and translation are also included.

https://doi.org/10.1017/nlp.2024.17 Published online by Cambridge University Press

in the dataset and have been miss-classified. Some slot labels have been labeled null tags in the slot filling task, primarily due to the scarcity of slot tags compared to null tags.

6. Conclusion

In this paper, we presented the first Persian public intent and slot filling dataset for task-oriented dialog systems, which consists of 500 samples for training and 962 samples for testing. This dataset is translated from English to Persian based on the ATIS dataset, and humans annotate its slots and intents labels. We evaluated the performance of different scenarios using rich-resource and low-resource data on ATIS datasets. To increase the reliability of the results of our different scenarios, we also repeated the experiments on the MASSIVE dataset. For both datasets, we consistently found that cross-lingual learning scenarios improve results compared to only training on limited amounts of data in a monolingual manner.

Future work will focus on adapting the model to deal with multi-intent scenarios. In addition, we will explore the possibility of training the model on the same scenarios to enable it to handle three or more languages simultaneously. Using data augmentation to overcome the problems of low-resource languages is another line of our future work.

References

Artetxe M., Labaka G. and Agirre E. (2017). Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada: Association for Computational Linguistics, pp. 451–462.

Castellucci G., Bellomaria V., Favalli A. and Romagnoli R. (2019). Multi-lingual intent detection and slot filling in a joint bert-based model, arXiv preprint arXiv: 1907.

Chen Q., Zhuo Z. and Wang W. (2019). Bert for joint intent classification and slot filling, arXiv preprint arXiv: 1902.10909.

Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave É., Ott M., Zettlemoyer L. and Stoyanov V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451.

FitzGerald J., Hench C., Peris C., Mackie S., Rottmann K., Sanchez A., Nash A., Urbach L., Kakarala V., Singh R., et al. (2022). Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages, arXiv preprint arXiv: 2204.08582.

Gritta M., Hu R. and Iacobacci I. (2022). Crossaligner & co: Zero-shot transfer methods for task-oriented cross-lingual natural language understanding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 4048–4061.

Gritta M. and Iacobacci I. (2021). Xeroalign: Zero-shot cross-lingual transformer alignment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP* **2021**(pages), 371–381.

Guo D., Tur G., Yih W.-t. and Zweig G. (2014). Joint semantic utterance classification and slot filling with recursive neural networks. In *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 554–559.

Hakkani-Tür D., Tür G., Celikyilmaz A., Chen Y.-N., Gao J., Deng L. and Wang Y.-Y. (2016). Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. In *Interspeech*, pp. 715–719.

Han S. C., Long S., Li H., Weld H. and Poon J. (2021). Bi-directional joint neural networks for intent classification and slot filling. In *Proceedings of Interspeech 2021*, pp. 4743–4747.

He K., Xu W. and Yan Y. (2020). Multi-level cross-lingual transfer learning with language shared and specific knowledge for spoken language understanding. *IEEE Access* **8**, 29407–29416.

Kenton J. D. M.-W. C. and Toutanova L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186.

Liu B. and Lane I. (2016). Attention-based recurrent neural network models for joint intent detection and slot filling. In *Interspeech*, pp. 685–689.

Liu Z., Shin J., Xu Y., Winata G. I., Xu P., Madotto A. and Fung P. (2019). 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)*, pp. 1297–1303.

McCann B., Bradbury J., Xiong C. and Socher R. (2017). Learned in translation: Contextualized word vectors. *Advances in Neural Information Processing Systems*, 30.

Mikolov T., Chen K., Corrado G. and Dean J. (2013). Efficient estimation of word representations in vector space, arXiv preprint arXiv: 1301.3781.

https://doi.org/10.1017/nlp.2024.17 Published online by Cambridge University Press

Pennington J., **Socher R. and Manning C. D.** (2014). Glove: Global vectors for word representation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Peters M. E., **Neumann M.**, **Iyyer M.**, **Gardner M.**, **Clark C.**, **Lee K. and Zettlemoyer L.** (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana: Association for Computational Linguistics, vol **1 (Long Papers)**, pp. 2227–2237.

Qin L., **Che W.**, **Li Y.**, **Wen H. and Liu T.** (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)*, pp. 2078–2087.

Qin L., **Ni M.**, **Zhang Y. and Che W.** (2021). CoSDA-ML: Multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP. In: *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 3853–3860.

Radford A., **Narasimhan K.**, **Salimans T. and Sutskever I.** (2018). Improving language understanding by generative pre-training.

Schuster S., **Gupta S.**, **Shah R. and Lewis M.** (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*, vol. **1 (Long and Short Papers)**, pp. 3795–3805.

Tang H., **Ji D. and Zhou Q.** (2020). End-to-end masked graph-based CRF for joint slot filling and intent detection. *Neurocomputing* **413**, 348–359.

Tur G., **Hakkani-Tür D. and Heck L.** (2010). What is left to be understood in ATIS?. In *2010 IEEE Spoken Language Technology Workshop*. IEEE, pp.19–24.

Upadhyay S., **Faruqui M.**, **Tür G.**, **Dilek H.-T. and Heck L.** (2018). (Almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6034–6038.

Wang Y., **Shen Y. and Jin H.** (2018). A bi-model based rnn semantic frame parsing model for intent detection and slot filling. In *Proceedings of NAACL-HLT*, pp. 309–314.

Xu W., **Haider B. and Mansour S.** (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)*, pp. 5052–5063.

Yang P., **Ji D.**, **Ai C. and Li B.** (2021). Aise: Attending to intent and slots explicitly for better spoken language understanding. *Knowledge-Based Systems* **211**, 106537.

Zhang C., **Li Y.**, **Du N.**, **Fan W. and Philip S. Y.** (2019). Joint slot filling and intent detection via capsule neural networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5259–5267.

Zhu S., **Cao R. and Yu K.** (2020). Dual learning for semi-supervised natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **28**, 1936–1947.

Cite this article: Zadkamali R, Momtazi S and Zeinali H. Intent detection and slot filling for Persian: Cross-lingual training for low-resource languages. *Natural Language Processing* https://doi.org/10.1017/nlp.2024.17