INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

FC-MTLF: A Fine- and Coarse-grained Multi-Task Learning Framework for Cross-Lingual Spoken Language Understanding

Xuxin Cheng, Wanshi Xu, Ziyu Yao, Zhihong Zhu, Yaowei Li, Hongxiang Li, Yuexian Zou[*]

ADSPLAB, School of ECE, Peking University, China
{chengxx, xwanshi, yaozy, zhihongzhu, ywl, lihongxiang}@stu.pku.edu.cn,
zouyx@pku.edu.cn

Abstract

Currently, zero-shot cross-lingual spoken language understanding (SLU) attracts increasing attention. Most of existing methods construct a mixed-language context via the code-switching approach. However, due to the different syntactic structures of each language, code-switching might fail to perform well and result in the loss of semantics. To address this issue, we propose a novel framework termed FC-MTLF, which applies a multi-task learning by introducing an auxiliary multilingual neural machine translation (NMT) task to compensate for the shortcomings of code-switching. In addition, we also adopt the curriculum learning strategy to further improve the performance. Experimental results show that our framework achieves the new state-of-the-art performance on the MultiATIS++ dataset. Further analysis verifies that our FC-MTLF can effectively transfer knowledge from source languages to target languages.

Index Terms: spoken language understanding, neural machine translation, multi-task learning, curriculum learning

1. Introduction

Spoken language understanding (SLU) [1, 2, 3, 4, 5] plays an important role in the task-oriented spoken dialog systems, with the aim of understanding user's current goal through constructing semantic frames. It usually includes two subtasks: intent detection and slot filling [6, 7, 8, 9, 10, 11]. With the advent of joint training models for intent detection and slot filling, SLU has gradually achieved remarkable success. However, most of existing SLU models rely on the large amounts of labeled training data, which greatly limits the scalability of the models to the low-resource languages with little or no training data. Zero-shot cross-lingual approaches have arisen to address this problem via transferring language-agnostic knowledge from high-resource (source) languages to low-resource (target) languages.

To this end, several works have been explored for zero-shot cross-lingual SLU. Multilingual BERT (mBERT) [12], a cross-lingual contextual pre-trained model from a large amount of multi-lingual corpus, has achieved the considerable performance for zero-shot cross-lingual SLU. [13] extends the idea to a multilingual code-switched setting, which aligns the source language to multiple target languages. [14] employs contrastive learning and additionally proposes a Local and Global component, which achieves the fine-grained cross-lingual transfer. LAJ-MCL [15] also applies contrastive learning to enhance the implicit alignment and feed to contrastive learning.

Most of the above-mentioned models that have achieved promising performance in zero-shot cross-lingual SLU adopt the code-switching method. However, there is a fatal problem in code-switching method. Due to different grammatical structures of each language, the target language chosen for the selected words according to the bilingual dictionary may not be very precise, which could result in semantic loss.

To address this problem, we propose a **F**ine- and **C**oarse-grained **M**ulti-**T**ask **L**earning **F**ramework (FC-MTLF) to perform explicit alignment. Specifically, we apply code-switching method to achieve the fine-grained explicit alignment and introduce neural machine translation (NMT) as an auxiliary task to achieve coarse-grained explicit alignment. During the process of machine translation, the semantic information is preserved between different languages, which could compensate for the semantic loss in code-switching to some extent. Furthermore, we introduce a curriculum learning strategy, which encourages the model to first focus on acquiring the ability to transfer the coarse-grained multilingual knowledge and then begin to gradually learn how to transfer fine-grained multilingual knowledge. Experiment results on the public benchmark MultiATIS++ [16] demonstrate that our FC-MTLF significantly surpasses the previous best zero-shot cross-lingual SLU models and model analysis further verifies the advantages of our method.

In summary, the contributions of our work are three-fold:

- We propose FC-MTLF for zero-shot cross-lingual SLU, which introduces NMT as an auxiliary task.
- We adopt the curriculum learning strategy to further improve the performance of the model.
- Our experiments on the MultiATIS++ benchmark show that FC-MTLF achieves the new state-of-the-art performance, with a relative improvement of 5.9% over the previous best model in average overall accuracy.

2. Method

We first describe the problem definition of traditional SLU task and zero-shot cross-lingual SLU task (§2.1). Then we introduce the three components of our framework, including Generic SLU Module (§2.2), Multiligual NMT Module (§2.3) and the curriculum learning strategy (§2.4), as shown in Figure 1.

2.1. Problem Formulation

SLU task contains two subtasks: intent detection and slot filling. Given an input utterance $x = (x_1, x_2, \dots, x_n)$, where n is the length of x. Intent detection can be formulated as a classification task which outputs the intent label $\boldsymbol{o}^I$. And slot filling is a sequence labeling task which maps each utterance x into a slot output sequence $\boldsymbol{o}^S = \left(o_1^S, o_2^S, \dots, o_n^S\right)$. Since the two tasks of intent detection and slot filling are highly related, it is common to train a single SLU model which is capable of handling these two tasks. We follow the formalism from [17], which is formulated as follows:

$$(\boldsymbol{o}^I, \boldsymbol{o}^S) = f(\mathbf{x}) \tag{1}$$

[*] Corresponding author.

10.21437/Interspeech.2023-46

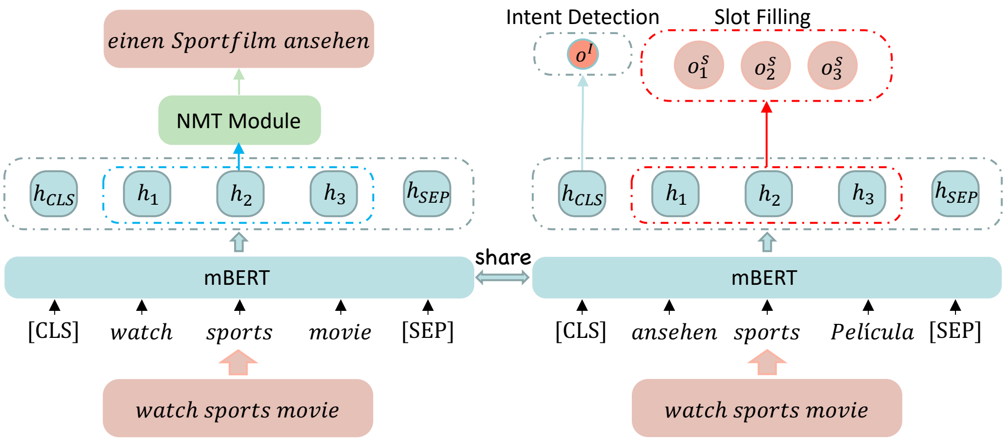

Figure 1: *The main architecture of FC-MTLF, which consists of two components: (1) a generic SLU module (*§2.2*); (2) a multilingual NMT module (*§2.3*). The mBERT models in SLU and NMT share the parameters. Moreover, we also apply a curriculum learning method (*§2.4*) to dynamically adjust the weights of these two tasks.*

where f is the trained model.

Zero-shot cross-lingual SLU task means that an SLU model is trained on the source language, e.g., English and directly adopted to the target language, e.g., German.

Specifically, given each instance $\mathbf{x}_{target}$ in the target language, the predicted intent and slot can be directly obtained by the model f which is trained on the source language:

$$\left(\boldsymbol{o}^I_{target}, \boldsymbol{o}^S_{target}\right) = f\left(\mathbf{x}_{target}\right) \tag{2}$$

where $target$ denotes the target language.

2.2. Generic SLU Module

Following previous work [12], given each input utterance $x = (x_1, x_2, ..., x_n)$, the input sequence can be constructed through adding specific tokens $\mathbf{x} = (\texttt{[CLS]}, x_1, x_2, ..., x_n, \texttt{[SEP]})$, where `[CLS]` denotes the special symbol for representing the whole sequence, and `[SEP]` can be used for separating the non-consecutive token sequences. We apply code-switching [13] to utilize the bilingual dictionaries [18] to generate multi-lingual code-switched data as the input of the model. Motivated by the success of pre-trained models in other tasks [19, 20], we get the representation of the sequence $\mathbf{H} = (\boldsymbol{h}_{\text{CLS}}, \boldsymbol{h}_1, \ldots, \boldsymbol{h}_n, \boldsymbol{h}_{\text{SEP}})$ by using mBERT [12] model. For the intent detection task, we input the utterence representation $\boldsymbol{h}_{\text{CLS}}$ to a classification layer to obtain the predicted intent:

$$\mathbf{o}^I = \text{softmax}\left(\boldsymbol{W}^I \boldsymbol{h}_{\text{CLS}} + \boldsymbol{b}^I\right) \tag{3}$$

where $\boldsymbol{W}^I$ and $\boldsymbol{b}^I$ denote the trainable matrices. For the slot filling task, we follow [21] and utilize the representation of the first sub-token as the whole word representation and utilize the hidden state to predict each slot:

$$\mathbf{o}^S_t = \text{softmax}\left(\boldsymbol{W}^s \boldsymbol{h}_t + \boldsymbol{b}^s\right) \tag{4}$$

where $\boldsymbol{h}_t$ denotes the representation of the first sub-token of word x_t, $\boldsymbol{W}^s$ and $\boldsymbol{b}^s$ denote the trainable matrices. Following previous work [17], the intent detection objective $\mathcal{L}_I$ and the slot filling objective $\mathcal{L}_S$ are as follows:

$$\mathcal{L}_I \triangleq -\sum_{i=1}^{n_I} \hat{\mathbf{y}}^I_i \log\left(\mathbf{o}^I_i\right) \tag{5}$$

$$\mathcal{L}_S \triangleq -\sum_{j=1}^{n}\sum_{i=1}^{n_S} \hat{\mathbf{y}}^{i,S}_j \log\left(\mathbf{o}^{i,S}_j\right) \tag{6}$$

where $\hat{\mathbf{y}}^I_i$ is the gold intent label, $\hat{\mathbf{y}}^{i,S}_j$ is the gold slot label for jth token, n_I is the number of intent labels, and n_S is the number of slot labels.

2.3. Multilingual NMT Module

To further enhance the coarse-grained explicit alignment, we introduce a multilingual neural machine translation task as the auxiliary task. Following [22, 23], we utilize the representation from mBERT as embeddings. Note that we do not apply code-switching method before NMT. Given the source language S, to promote the knowledge transfer and improve the robustness to inevitable label noise [24], we randomly choose an another language T as the target language, the training loss is the cross entropy loss $\mathcal{L}_{CE}$:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{|x^T|} \log p_\theta(x^T_i | x^T_{<i}, x^S) \tag{7}$$

where x^S denotes an utterance in the source language, x^T denotes the corresponding utterance in the target language, θ denotes the parameter of the NMT module.

2.4. Curriculum Learning

Motivated by the success of curriculum learning [25, 26], we introduce a curriculum learning strategy. The final loss $\mathcal{L}$ of the multi-task learning framework is:

$$\mathcal{L} = \lambda(\alpha\mathcal{L}_I + \beta\mathcal{L}_S) + (1 - \lambda)\mathcal{L}_{CE} \tag{8}$$

$$\lambda = \frac{k}{K} \tag{9}$$

where α and β are hyper-parameters, k is the current number of updates, K is the total number of updates, λ is determined by k and K and dynamic changes during the training process.

3. Experiments

3.1. Datasets and Metrics

We conduct the experiments on the cross-lingual SLU benchmark dataset, MultiATIS++[1] [16], which contains 18 intents and 84 slots for each language. MultiATIS++ further extends Multilingual ATIS by adding the human-translated data for six additional languages, including Spanish (es), German (de), Chinese (zh), Japanese (ja), Portuguese (pt), and French (fr), to the

[1] `https://github.com/amazon-science/multiatis`

Table 1: *Experiment results on MultiATIS++. We report both individual and average (AVG) intent detection accuracy, slot filling F1 score and overall accuracy. Results with "*" are taken from the corresponding published paper, results with † are cited from [14], and results with ‡ are cited from [15]. '–' denotes missing results from the published work.*

Intent Accuracy	en	de	es	fr	hi	ja	pt	tr	zh	**AVG**
mBERT* [16]	-	95.27	96.35	95.92	80.96	79.42	94.96	69.59	86.27	-
mBERT† [12]	98.54	95.40	96.30	94.31	82.41	76.18	94.95	75.10	82.53	88.42
ZSJoint‡ [27]	98.54	90.48	93.28	94.51	77.15	76.59	94.62	73.29	84.55	87.00
Ensemble-Net* [28]	90.26	92.50	96.64	95.18	77.88	77.04	95.30	75.04	84.99	87.20
CoSDA† [13]	95.74	94.06	92.29	77.04	82.75	73.25	93.05	80.42	78.95	87.32
GL-CLeF* [14]	98.77	97.53	97.05	97.72	86.00	82.84	96.08	83.92	87.68	91.95
LAJ-MCL* [15]	98.77	98.10	98.10	98.77	84.54	81.86	97.09	85.45	89.03	92.41
FC-MTLF	**98.97**	**98.21**	**98.36**	**99.01**	**86.72**	**82.95**	**97.34**	**86.02**	**89.53**	**93.01**
Slot F1	en	de	es	fr	hi	ja	pt	tr	zh	**AVG**
Ensemble-Net* [28]	85.05	82.75	77.56	76.19	14.14	9.44	74.00	45.63	37.29	55.78
mBERT* [16]	-	82.61	74.98	75.71	31.21	35.75	74.05	23.75	62.27	-
mBERT† [12]	95.11	80.11	78.22	82.25	26.71	25.40	72.37	41.49	53.22	61.66
ZSJoint‡ [27]	95.20	74.79	76.52	74.25	52.73	70.10	72.56	29.66	66.91	68.08
CoSDA† [13]	92.29	81.37	76.94	79.36	64.06	66.62	75.05	48.77	77.32	73.47
GL-CLeF* [14]	95.39	86.30	85.22	84.31	70.34	73.12	81.83	65.85	77.61	80.00
LAJ-MCL* [15]	96.02	86.59	83.03	82.11	61.04	68.52	81.49	65.20	82.00	78.23
FC-MTLF	**96.21**	**86.87**	**85.66**	**84.62**	**73.18**	**74.24**	**82.68**	**68.22**	**83.16**	**81.65**
Overall Accuracy	en	de	es	fr	hi	ja	pt	tr	zh	**AVG**
AR-S2S-PTR* [29]	86.83	34.00	40.72	17.22	7.45	10.04	33.38	–	23.74	-
IT-S2S-PTR* [30]	87.23	39.46	50.06	46.78	11.42	12.60	39.30	–	28.72	-
mBERT† [12]	87.12	52.69	52.02	37.29	4.92	7.11	43.49	4.33	18.58	36.29
ZSJoint‡ [27]	87.23	41.43	44.46	43.67	16.01	33.59	43.90	1.12	30.80	38.02
CoSDA† [13]	77.04	57.06	46.62	50.06	26.20	28.89	48.77	15.24	46.36	44.03
GL-CLeF* [14]	88.02	66.03	59.53	57.02	34.83	41.42	60.43	28.95	50.62	54.09
LAJ-MCL* [15]	89.81	67.75	59.13	57.56	23.29	29.34	61.93	28.95	54.76	52.50
FC-MTLF	**91.58**	**69.54**	**61.43**	**59.62**	**36.86**	**44.64**	**64.55**	**30.86**	**56.52**	**57.29**

initial Hindi (hi) and Turkish (tr) languages. Following previous works [17], we evaluate intent prediction performance using accuracy, slot filling performance using F1 score, and sentence-level semantic frame parsing using overall accuracy.

3.2. Implementation Details

We use the base case multilingual BERT (mBERT), which has $N = 12$ attention heads and $M = 12$ transformer blocks. In Eq.8, α is set to 0.9 and β is set to 0.1. The total number of updates K is 30k. We utilize Adam optimizer [31] with $\beta_1 = 0.9, \beta_2 = 0.98$ to optimize parameters in our model. All experiments are conducted at an Nvidia Tesla-V100. The training process lasts several hours.

3.3. Baselines

We compare our model with 8 baselines: (1) `mBERT`: mBERT[2] follows the same model architecture as BERT [12], but instead of training only on monolingual English data, it is trained on the Wikipedia pages of 104 languages with a shared word piece vocabulary, allowing the model to share embeddings across languages; (2) `AR-S2S-PTR`: [29] proposes a unified sequence-to-sequence models with the pointer generator network for cross-lingual SLU; (3) `IT-S2S-PTR`: [30] proposes a non-autoregressive parser based on insertion transformer, which speeds up the decoding progress; (4) `Ensemble-Net`: [28] proposes an effective zero-shot cross-lingual SLU model, whose predictions are the majority voting results of 8 independent models, each separately trained on a single source language; (5) `ZSJoint`: [27] proposes a zero-shot SLU model, which is trained on the en training set and directly applied to the test sets of target languages; (6) `CoSDA`: [13] proposes a data augmentation framework to generate multi-lingual code-switching data to fine-tune mBERT, which encourages the model to align representations from the source and multiple target languages; (7) `GL-CLeF`: [14] introduces a contrastive learning framework to explicitly align representations across languages for zero-shot cross-lingual SLU; (8) `LAJ-MCL`: [15] proposes a multi-level contrastive learning framework for zero-shot cross-lingual SLU.

3.4. Main Results

From the experimental results on MultiATIS++ dataset shown in Table 1, we have the following observations:

(1) CoSDA, GL-CLeF and LAJ-MCL all apply code-switching, we can find they outperform the model that does not use this method. The reason is that code-switching achieves an implicit alignment, which can align the representations to some extent. In addition, our method further improves the performance compared to these three models and obtains a relative improvement of 5.9% over the previous best model in average overall accuracy. This is because our method introduces NMT as an auxiliary task, which makes better use of the connections between different languages.

(2) FC-MTLF achieves significant and consistent improve-

[2] `https://github.com/google-research/bert/blob/master/multilingual.md`

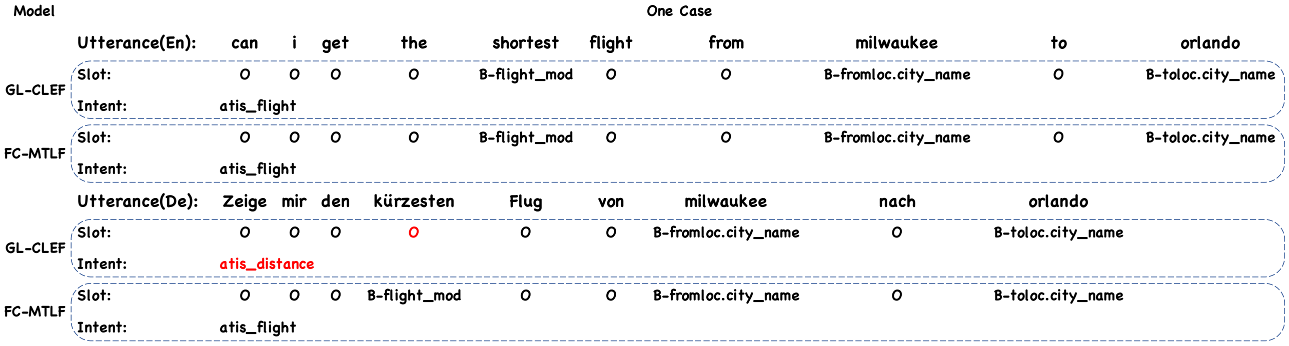
Model
One Case
Utterance(En): can i get the shortest flight from milwaukee to orlando
GL-CLEF
Slot: O O O O B-flight_mod O O B-fromloc.city_name O B-toloc.city_name
Intent: atis_flight
FC-MTLF
Slot: O O O O B-flight_mod O O B-fromloc.city_name O B-toloc.city_name
Intent: atis_flight
Utterance(De): Zeige mir den kürzesten Flug von milwaukee nach orlando
GL-CLEF
Slot: O O O O O O B-fromloc.city_name O B-toloc.city_name
Intent: atis_distance
FC-MTLF
Slot: O O O B-flight_mod O O B-fromloc.city_name O B-toloc.city_name
Intent: atis_flight

Figure 2: *A case study of our framework compared to GL-CLEF [14]. Intents and slots in red refer to those that are predicted incorrectly.*

Table 2: *Ablation results on the MultiATIS++ dataset.*

Models	Intent	Slot	Overall
FC-MTLF	**93.01**	**81.65**	**57.29**
w/o NMT	87.32(↓5.69)	73.47(↓8.18)	44.03(↓13.26)
More Parameters	88.24(↓4.77)	74.86(↓6.79)	46.52(↓10.77)
w/o CL	92.58(↓0.43)	81.06(↓0.59)	56.61(↓0.68)

ments in all subtasks. Moreover, compared to high-resource languages, it obtains larger improvement on low-resource languages. We can find that there is relatively little training data for Hindi and Turkish in MultiATIS++ compared to other languages and the improvement in these two languages is much greater than that of other high-resource languages. Zero-shot cross-lingual SLU task is originally designed to solve the problem of low-resource languages with little training data. This result demonstrates that our method could effectively transfer knowledge from source languages to target languages, which further verifies the superiority of our framework.

3.5. Model Analysis

3.5.1. Effect of Neural Machine Translation

To demonstrate the effectiveness of neural machine translation, we remove it and refer it to *w/o NMT* in Table 2. We can observe that after we remove the NMT module, the intent accuracy of MixATIS++ drops by 5.69%. In addition, the overall accuracy drops by 13.26%, which is an even more significant decrease. This illustrates the importance of the NMT module in our model, which enhances the coarse-grained explicit alignment between different languages.

3.5.2. Effect of More Parameters

Following the previous works [6, 7], to verify whether the increased parameters of FC-MTLF contribute to the final higher performance, we add an LSTM layer after the last layer of mBERT and refer it to *More Parameters*. The results in Table 2 demonstrate that our framework outperforms mBERT with more parameters in intent accuracy, slot F1 and overall accuracy, which verifies that the improvement indeed comes from the fine-grained and coarse-grained multi-task learning framework rather than the involved parameters.

3.5.3. Effect of Curriculum Learning

To demonstrate the effectiveness of curriculum learning, we design a variant termed *w/o CL* and the results are shown in Table 2. We remove the curriculum strategy and utilize static coefficients. After several experiments, the model works best when λ in Eq.9 is set to 0.6. So we set λ to 0.6 for this ablation experiment. We can clearly find that curriculum learning strategy is very effective in improving the performance of the model, which encourages the model to initially concentrate on obtaining proficiency in transferring coarse-grained multilingual knowledge and then gradually learns how to transfer fine-grained multilingual knowledge.

3.6. Case Study

To further demonstrate the advancement of our model relative to previous models working on the zero-shot cross-lingual SLU task, we provide a case study in which we illustrate the effects of our model on different languages. We use English and German as examples. As shown in Figure 2, we can find that the prediction results of both GL-CLEF and FC-MTLF are correct for English. For the utterance in German with the same meaning, the predicted slots and intents of GL-CLEF are inaccuracy, while FC-MTLF could predict correctly. This suggests that our model achieves better cross-lingual transfer with higher generalizability through the introduced NMT module.

4. Conclusions

In this paper, we propose FC-MTLF, a novel multi-task learning framework for zero-shot cross-lingual spoken language understanding (SLU), which utilizes code-switching for coarse-grained alignment and NMT for fine-grained alignment. In addition, we apply a curriculum learning strategy to further improve the performance. Experiments on MultiATIS++ dataset show that FC-MTLF achieve a new state-of-the-art performance. And model analysis indicates that FC-MTLF successfully transfers knowledge from source languages to target languages. Future work will focus on how to further achieve explicit alignment to improve the performance.

5. Acknowledgements

We thank all the anonymous reviewers and program chairs for their valuable and insightful feedback. This paper was partially supported by Shenzhen Science & Technology Research Program (No:GXWD20201231165807007-20200814115301001) and NSFC (No: 62176008) .

6. References

[1] G. Tur and R. De Mori, *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons, 2011.

[2] S. Young, M. Gašić, B. Thomson, and J. D. Williams, “Pomdp-based statistical spoken dialog systems: A review,” *Proceedings of the IEEE*, 2013.

[3] P. Xu and R. Sarikaya, “Convolutional neural network based triangular crf for joint intent detection and slot filling,” in *2013 ieee workshop on automatic speech recognition and understanding*, 2013.

[4] B. Kim, S. Ryu, and G. G. Lee, “Two-stage multi-intent detection for spoken language understanding,” *Multimedia Tools and Applications*, 2017.

[5] R. Gangadharaiah and B. Narayanaswamy, “Joint multiple intent detection and slot labeling for goal-oriented dialog,” in *Proc. of NAACL*, 2019.

[6] L. Qin, X. Xu, W. Che, and T. Liu, “AGIF: An adaptive graph-interactive framework for joint multiple intent detection and slot filling,” in *Proc. of EMNLP Findings*, 2020.

[7] L. Qin, F. Wei, T. Xie, X. Xu, W. Che, and T. Liu, “GL-GIN: Fast and accurate non-autoregressive model for joint multiple intent detection and slot filling,” in *Proc. of ACL*, 2021.

[8] B. Xing and I. W. Tsang, “Group is better than individual: Exploiting label topologies and label relations for joint multiple intent detection and slot filling,” *ArXiv preprint*, 2022.

[9] X. Cheng, B. Cao, Q. Ye, Z. Zhu, H. Li, and Y. Zou, “Ml-lmcl: Mutual learning and large-margin contrastive learning for improving asr robustness in spoken language understanding,” in *Proc. of ACL Findings*, 2023.

[10] Z. Zhu, X. Cheng, Z. Huang, D. Chen, and Y. Zou, “Towards unified spoken language understanding decoding via label-aware compact linguistics representations,” in *Proc. of ACL Findings*, 2023.

[11] Z. Zhu, W. Xu, X. Cheng, T. Song, and Y. Zou, “A dynamic graph interactive framework with label-semantic injection for spoken language understanding,” in *Proc. of ICASSP*, 2023.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. of NAACL*, 2019.

[13] L. Qin, M. Ni, Y. Zhang, and W. Che, “Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP,” in *Proc. of IJCAI*, 2020.

[14] L. Qin, Q. Chen, T. Xie, Q. Li, J.-G. Lou, W. Che, and M.-Y. Kan, “GL-CLeF: A global–local contrastive learning framework for cross-lingual spoken language understanding,” in *Proc. of ACL*, 2022.

[15] S. Liang, L. Shou, J. Pei, M. Gong, W. Zuo, X. Zuo, and D. Jiang, “Label-aware multi-level contrastive learning for cross-lingual spoken language understanding,” in *Proc. of EMNLP*, 2022.

[16] W. Xu, B. Haider, and S. Mansour, “End-to-end slot alignment and recognition for cross-lingual NLU,” in *Proc. of EMNLP*, 2020.

[17] C.-W. Goo, G. Gao, Y.-K. Hsu, C.-L. Huo, T.-C. Chen, K.-W. Hsu, and Y.-N. Chen, “Slot-gated modeling for joint slot filling and intent prediction,” in *Proc. of NAACL*, 2018.

[18] G. Lample, A. Conneau, M. Ranzato, L. Denoyer, and H. Jégou, “Word translation without parallel data,” in *Proc. of ICLR*, 2018.

[19] X. Cheng, Q. Dong, F. Yue, T. Ko, M. Wang, and Y. Zou, “M 3 st: Mix at three levels for speech translation,” in *Proc. of ICASSP*, 2023.

[20] Z. Zhu, X. Cheng, D. Chen, Z. Huang, H. Li, and Y. Zou, “Mix before Align: Towards Zero-shot Cross-lingual Sentiment Analysis via Soft-Mix and Multi-View Learning,” in *Proc. of Interspeech*, 2023.

[21] Y. Wang, W. Che, J. Guo, Y. Liu, and T. Liu, “Cross-lingual BERT transformation for zero-shot dependency parsing,” in *Proc. of EMNLP*, 2019.

[22] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” in *Proc. of NAACL*, 2018.

[23] J. Zhu, Y. Xia, L. Wu, D. He, T. Qin, W. Zhou, H. Li, and T. Liu, “Incorporating BERT into neural machine translation,” in *Proc. of ICLR*, 2020.

[24] X. Cheng, Z. Zhu, H. Li, Y. Li, and Y. Zou, “Ssvmr: Saliency-based self-training for video-music retrieval,” in *Proc. of ICASSP*, 2023.

[25] J. Guo, X. Tan, L. Xu, T. Qin, E. Chen, and T. Liu, “Fine-tuning by curriculum learning for non-autoregressive neural machine translation,” in *Proc. of AAAI*, 2020.

[26] L. Qian, H. Zhou, Y. Bao, M. Wang, L. Qiu, W. Zhang, Y. Yu, and L. Li, “Glancing transformer for non-autoregressive neural machine translation,” in *Proc. of ACL*, 2021.

[27] Q. Chen, Z. Zhuo, and W. Wang, “Bert for joint intent classification and slot filling,” *ArXiv preprint*, 2019.

[28] E. Razumovskaia, G. Glavas, O. Majewska, E. M. Ponti, A. Korhonen, and I. Vulic, “Crossing the conversational chasm: A primer on natural language processing for multilingual task-oriented dialogue systems,” *Journal of Artificial Intelligence Research*, 2022.

[29] S. Rongali, L. Soldaini, E. Monti, and W. Hamza, “Don’t parse, generate! A sequence to sequence architecture for task-oriented semantic parsing,” in *Proc. of WWW*, 2020.

[30] Q. Zhu, H. Khan, S. Soltan, S. Rawls, and W. Hamza, “Don’t parse, insert: Multilingual semantic parsing with insertion based decoding,” in *Proc. of CoNLL*, 2020.

[31] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. of ICLR*, 2015.