INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

DiffSLU: Knowledge Distillation Based Diffusion Model for Cross-Lingual Spoken Language Understanding

*Tianjun Mao, Chenghong Zhang**

School of Management, Fudan University, Shanghai, China

tjmao22@m.fudan.edu.cn, chzhang@fudan.edu.cn

Abstract

Spoken language understanding (SLU) has achieved great success in high-resource languages, but it still remains challenging in the low-resource languages due to the scarcity of labeled training data. Hence, there is an increasing interest in zero-shot cross-lingual SLU. SLU typically has two subtasks, including intent detection and slot filling. Slots and intent in the same utterance are correlated, thus it is beneficial to achieve mutual guidance between them. In this paper, we propose a novel cross-lingual SLU framework termed DiffSLU, which leverages powerful diffusion model to enhance the mutual guidance. In addition, we also utilize knowledge distillation to facilitate knowledge transfer. Experimental results demonstrate that our DiffSLU can improve the performance compared with the strong baselines and achieves the new state-of-the-art performance on MultiATIS++ dataset, obtaining a relative improvement of 3.1% over the previous best model in overall accuracy.

Index Terms: spoken language understanding, zero-shot, diffusion model, knowledge distillation

1. Introduction

Task-oriented dialogue systems rely on spoken language understanding (SLU) [1, 2, 3, 4], which aims to extract semantic components from queries. Typically, SLU comprises of two subtasks, including intent detection and slot filling [5, 6]. The emergence of deep neural network techniques has led to remarkable accomplishments in the field of SLU. Nonetheless, most of them demand extensive labeled training data, which restricts the performance on the languages with scarce or no training data, thereby limiting their scalability. To tackle this issue, zero-shot cross-lingual SLU [7, 8] has garnered lots of attention as it uses labeled data in high-resource languages to transfer knowledge from trained models to low-resource target languages.

Recently, numerous studies are carried out to achieve zero-shot cross-lingual SLU, including the utilization of Multilingual BERT (mBERT) [9], a contextual pre-trained model trained on a corpus of multiple languages, which shows significant progress in achieving zero-shot cross-lingual SLU. In a code-switched setting, [10] expands on the notion by aligning the source language with several target languages, which simply utilizes bilingual dictionaries to randomly select some words in the utterance to be replaced by the translation of the words in other languages. [7] and [8] apply contrastive learning to achieve explicit alignment to further improve the performance. However, most of the previous works neglect to implement the mutual guidance between intent and slots, which is beneficial to the performance of the SLU model. In this paper, we propose DiffSLU to leverage powerful diffusion model to achieve the mutual guidance.

Continuous diffusion models are first successfully applied in image generation [11, 12, 13]. Recently, diffusion models are applied to sequence learning and controllable text generation. Diffusion-LM [14] is the first diffusion model for sequence generate, which demonstrates the superiority of diffusion models in generating sequences. The improvements to diffusion-based sequence generative models can be broadly categorized into three lines. The first line introduces some novel components, such as the partial diffusion process proposed by [15], self-conditioning techniques introduced by [16], and the adaptive noise schedule proposed by [17]. The second line applies diffusion models to the latent space of pre-trained language models [18]. The third line attempts to combine conventional practices in discrete token prediction with diffusion models. For instance, [19, 20, 21] incorporate the cross-entropy objectives in training.

DiffSLU consists of two components with the same architecture. The training process includes forward process with partial noising and reverse process with conditional denoising [15]. Specifically, we first apply code-switching method [10] to generate multi-lingual code-switched utterance based on the original utterance. We concatenate the representation of the original utterance, the one-hot encoding of the slot label, and the one-hot encoding of the intent label as the input of the first diffusion model. Similarly, we concatenate the representation of the code-switched utterance and the two one-hot encoding as the input of the second diffusion model. Through concatenating the two one-hot encoding, our method achieves the mutual guidance between these two subtasks. Following [15], we only impose noising on the concatenation of the two one-hot encoding. In addition, we apply knowledge distillation in the reverse process to transfer knowledge from the first diffusion model to the second diffusion model, which could further improve the performance. In the inference process, only the second diffusion model is utilized. Experiment results on the public benchmark dataset MultiATIS++ [22] show that DiffSLU significantly outperforms the previous best zero-shot cross-lingual SLU models and analysis further verifies the advantages of our method.

In summary, the contributions of this paper are as follows:

- To the best of our knowledge, we are the first to apply diffusion models for zero-shot cross-lingual SLU.
- We apply knowledge distillation to further improve the performance of the model.
- Experiments show that DiffSLU achieves a new state-of-the-art performance, obtaining an improvement of 3.1% over the previous best model in average overall accuracy.

* Corresponding author.

10.21437/Interspeech.2023-699

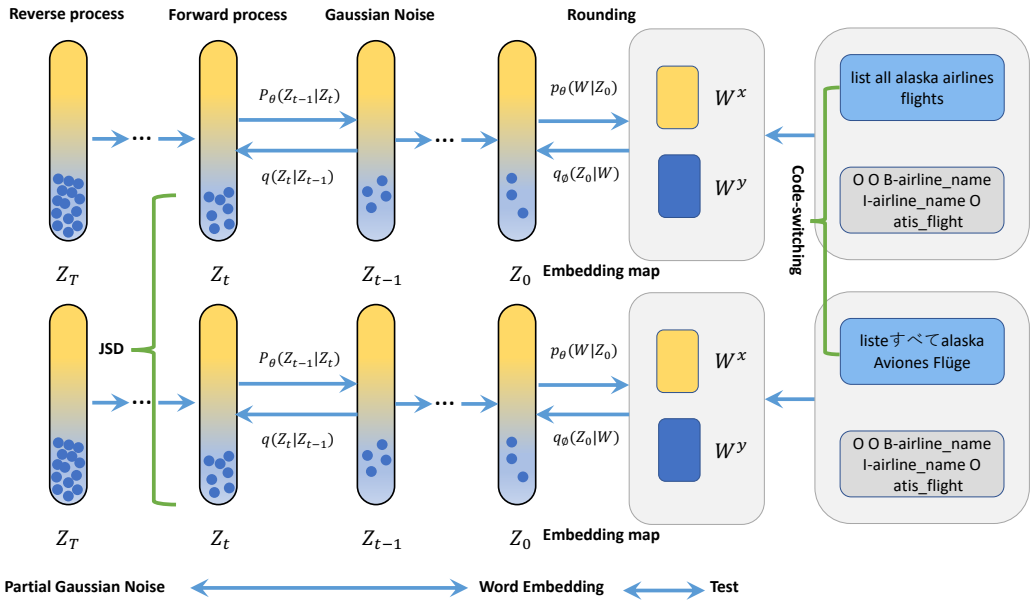

Figure 1: *The overview of our DiffSLU. Two models with the same architecture are trained on the original utterance and code-switched utterance, respectively. Jensen-Shannon Divergence (JSD) is applied to transfer knowledge.*

2. Method

In this section, we first describe the background (§2.1) of cross-lingual SLU. Then we introduce the main architecture of DiffSLU. Finally we introduce the overall training objective (§2.3). The overview of our method is shown in Figure 1.

2.1. Background

Intent detection and slot filling are two subtasks of SLU. Given an input utterance $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, where n is the length of $\mathbf{x}$. Intent detection is a classification task which predicts the intent $\boldsymbol{o}^I$. Slot filling is a sequence labeling task which maps each utterance $\mathbf{x}$ into a slot sequence $\boldsymbol{o}^S = (o_1^S, o_2^S, \ldots, o_n^S)$. Training a single model that can handle both tasks of intent detection and slot filling is a common practice as they are closely interconnected. Following previous work [23], the formalism is formulated as follows:

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

where f is the trained model.

The zero-shot cross-lingual SLU task involves training an SLU model in a source language and then adapting it directly to target languages without additional training. Specifically, given each instance $\mathbf{x}_{tgt}$ in the target language, the predicted intent and slot can be directly obtained by the SLU model f which is trained on the source language:

$$\left(\boldsymbol{o}_{tgt}^I, \boldsymbol{o}_{tgt}^S\right) = f\left(\mathbf{x}_{tgt}\right) \tag{2}$$

where tgt denotes the target language.

2.2. Main Architecture

Inspired by the accomplishment of pre-trained models in other tasks [24, 25], we follow [7] to obtain the representation $\mathbf{H}$ of the utterance $\mathbf{x}$ by using mBERT [9] model:

$$\mathbf{H} = (\boldsymbol{h}_{\texttt{CLS}}, \boldsymbol{h}_1, \ldots, \boldsymbol{h}_n, \boldsymbol{h}_{\texttt{SEP}}) \tag{3}$$

where `[CLS]` denotes the special symbol for representing the whole sequence, and `[SEP]` can be utilized for separating non-consecutive token sequences.

For intent detection, we input the utterance representation $\boldsymbol{h}_{\texttt{CLS}}$ to a classification layer to obtain the predicted intent:

$$\boldsymbol{o}^I = \text{softmax}\left(\boldsymbol{W}^I \boldsymbol{h}_{\texttt{CLS}} + \boldsymbol{b}^I\right) \tag{4}$$

where $\boldsymbol{W}^I$ and $\boldsymbol{b}^I$ denote the trainable matrices.

For slot filling, we follow [26] to utilize the representation of the first sub-token as the whole word representation and use the hidden state to predict each slot:

$$\boldsymbol{o}_t^S = \text{softmax}\left(\boldsymbol{W}^s \boldsymbol{h}_t + \boldsymbol{b}^s\right) \tag{5}$$

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.

We employ code-switching [10] to leverage the bilingual dictionaries [27] to generate the multi-lingual code-switched utterance $\mathbf{x}'$. We denote the representation of the original utterance $\mathbf{x}$ as $\mathbf{H}^o$ and the representation of the code-switched utterance $\mathbf{x}'$ as $\mathbf{H}^c$. Motivated by [15], we concatenate the representation $\mathbf{H}^o$, the one-hot encoding $\hat{\mathbf{y}}^S$ of the slot label, and the one-hot encoding $\hat{\mathbf{y}}^I$ of the intent label as the input of the first diffusion model. Similarly, we concatenate the representation $\mathbf{H}^c$ and the two one-hot encoding $\hat{\mathbf{y}}^S$ and $\hat{\mathbf{y}}^I$ as the input of the second diffusion model. In the forward process, we only impose noising on the concatenation $\hat{\mathbf{y}}$ of $\hat{\mathbf{y}}^S$ and $\hat{\mathbf{y}}^I$. We denote the concatenation of $\mathbf{H}^o$ and $\hat{\mathbf{y}}$ as $\mathbf{E}^1$ and the concatenation of $\mathbf{H}^c$ and $\hat{\mathbf{y}}$ as $\mathbf{E}^2$. We pair-wisely transform $\mathbf{E}^1$ and $\mathbf{E}^2$ into continuous space $\mathbf{z}_0^1$ and $\mathbf{z}_0^2$. The loss $\mathcal{L}_{dif}$ of the diffusion model is formulated as follows:

$$\begin{aligned}\mathcal{L}_{dif} = &\sum_{t=2}^{T} ||\mathbf{z}_0^1 - f_\theta^1(\mathbf{z}_t, t)||^2 + \sum_{t=2}^{T} ||\mathbf{z}_0^2 - f_\theta^2(\mathbf{z}_t, t)||^2 \\ &+ ||\mathbf{E}^1 - f_\theta^1(\mathbf{z}_1, 1)||^2 + ||\mathbf{E}^2 - f_\theta^2(\mathbf{z}_1, 1)||^2\end{aligned} \tag{6}$$

where $f_\theta^1(\mathbf{z}_t, t)$ is denoted the fractions of recovered $\mathbf{z}_0^1$ corresponding to $\hat{\mathbf{y}}$ and $f_\theta^2(\mathbf{z}_t, t)$ is denoted the fractions of recovered $\mathbf{z}_0^2$ corresponding to $\hat{\mathbf{y}}$.

Besides, we also apply knowledge distillation to transfer the knowledge and enhance the robustness to the label noise [28]. We randomly choose a time t and utilize the Jensen-Shannon Divergence (JSD) between $f_\theta^1(\mathbf{z}_t, t)$ of the first model and $f_\theta^2(\mathbf{z}_t, t)$ of the seconde model:

$$\mathcal{L}_{kd} = JSD(f_\theta^1(\mathbf{z}_t, t), f_\theta^2(\mathbf{z}_t, t)) \tag{7}$$

2.3. Training Objective

Following previous work [23], the intent detection objective $\mathcal{L}_I$ and the slot filling objective $\mathcal{L}_S$ are formulated as follows:

$$\mathcal{L}_I \triangleq -\sum_{i=1}^{n_I} \hat{\mathbf{y}}_i^I \log\left(\mathbf{o}_i^I\right) \tag{8}$$

$$\mathcal{L}_S \triangleq -\sum_{j=1}^{n}\sum_{i=1}^{n_S} \hat{\mathbf{y}}_j^{i,S} \log\left(\mathbf{o}_j^{i,S}\right) \tag{9}$$

where $\hat{\mathbf{y}}_i^I$ is the gold intent label, $\hat{\mathbf{y}}_j^{i,S}$ 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.

The final training objective is as follow:

$$\mathcal{L} = \alpha\mathcal{L}_I + \beta\mathcal{L}_S + \lambda\mathcal{L}_{dif} + \gamma\mathcal{L}_{kd} \tag{10}$$

where α, β, λ, γ are the hyper-parameters.

3. Experiments

3.1. Datasets and Metrics

All the experiments are conducted on the cross-lingual SLU benchmark dataset, MultiATIS++[1][22], which contains 18 intents and 84 slots for each language. Human-translated data for six languages including Spanish (es), German (de), Chinese (zh), Japanese (ja), Portuguese (pt), French (fr) are added to Multilingual ATIS which has Hindi (hi) and Turkish (tr). The statistics of MultiATIS++ dataset are shown in Table 1.

Table 1: *Statistics of MultiATIS++*

Language	Utterances			Intent types	Slot types
	train	valid	test		
hi	1440	160	893	17	75
tr	578	60	715	17	71
others	4488	490	893	18	84

Following the previous works [7, 8, 23], we utilize accuracy to evaluate the intent prediction performance, F1 score to evaluate the slot filling performance, and overall accuracy to evaluate the sentence-level semantic frame parsing. Overall accuracy represents whether all metrics including intent and slots in the utterance are correctly predicted.

3.2. Implementation Details

The mBERT model we utilized has $N = 12$ attention heads and $M = 12$ transformer blocks. Following previous work [7], we select the best hyperparameters by searching a combination of batch size, learning rate with the following ranges: learning rate $\{2 \times 10^{-7}, 5 \times 10^{-7}, 1 \times 10^{-6}, 2 \times 10^{-6}, 5 \times 10^{-6}, 6 \times 10^{-6}, 5 \times 10^{-5}, 5 \times 10^{-4}\}$ and batch size $\{4, 8, 16, 32\}$. α, β, λ, γ are set to 0.9, 0.1, 0.8 and 0.2 in Eq.10, respectively. We use Adam optimizer [33] with $\beta_1 = 0.9, \beta_2 = 0.98$ to optimize the parameters. For all the experiments, we select the model that performs the best on the `dev` set in terms of overall accuracy and evaluate it on the `test` set. The training process will early-stop if the loss on the `dev` set did not decrease for 5 epochs. All experiments are conducted at an Nvidia Tesla-V100 GPU. The training process lasts several hours.

3.3. Baselines

We compare our model to the following baselines:

- `mBERT`: mBERT[2] follows the same model architecture and training procedure as BERT [9], 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;
- `AR-S2S-PTR`: a unified sequence-to-sequence models with the pointer generator network proposed by [31] for zero-shot cross-lingual SLU;
- `IT-S2S-PTR`: a non-autoregressive model based on the insertion transformer proposed by [32], which speeds up the decoding progress of zero-shot cross-lingual SLU;
- `Ensemble-Net`: [30] 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;
- `ZSJoint`: [29] proposes a zero-shot SLU model, which is trained on the en training set and directly applied to the test sets of target languages.
- `CoSDA`: [10] 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.
- `GL-CLeF`: [7] introduces a contrastive learning framework to explicitly align representations across languages for zero-shot cross-lingual SLU.
- `LAJ-MCL`: [8] proposes a multi-level contrastive learning framework for zero-shot cross-lingual SLU.

3.4. Main Results

The performance comparison of DiffSLU and baselines are shown in Table 2, from which we have the following observations: (1) The models which applies code-switching method including CoSDA, GL-CLeF and LAJ-MCL outperform the models which do not use this method. This is because code-switching produces an implicit alignment, thereby aligning the representations to some degree. (2) Moreover, DiffSLU further improves the performance and obtains a relative improvement of 3.1% over the previous best model in terms of average overall accuracy. The reason is that our method enhance the mutual guidance between intent and slots, which is helpful to further improve the performance of the model.

3.5. Model Analysis

3.5.1. Effect of Diffusion Module

We remove the diffusion module and refer it to *w/o diffusion* in Table 3 to verify the effectiveness. It is obvious that after we remove the diffusion module, the intent accuracy of MixATIS++ dataset drops by 5.58%. Moreover, the overall accuracy also

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

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

Table 2: *Experiment results on the MultiATIS++ dataset. '–' denotes missing results from the published work.*

Intent (Acc)	en	de	es	fr	hi	ja	pt	tr	zh	**AVG**
mBERT [22]	-	95.27	96.35	95.92	80.96	79.42	94.96	69.59	86.27	-
mBERT [9]	98.54	95.40	96.30	94.31	82.41	76.18	94.95	75.10	82.53	88.42
ZSJoint [29]	98.54	90.48	93.28	94.51	77.15	76.59	94.62	73.29	84.55	87.00
Ensemble-Net [30]	90.26	92.50	96.64	95.18	77.88	77.04	95.30	75.04	84.99	87.20
CoSDA [10]	95.74	94.06	92.29	77.04	82.75	73.25	93.05	80.42	78.95	87.32
GL-CLEF [7]	98.77	97.53	97.05	97.72	86.00	82.84	96.08	83.92	87.68	91.95
LAJ-MCL [8]	98.77	98.10	98.10	98.77	84.54	81.86	97.09	85.45	89.03	92.41
DiffSLU	**98.86**	**98.17**	**98.21**	**98.93**	**86.66**	**82.65**	**97.21**	**85.98**	**89.46**	**92.90**
Slot (F1)	en	de	es	fr	hi	ja	pt	tr	zh	**AVG**
Ensemble-Net [30]	85.05	82.75	77.56	76.19	14.14	9.44	74.00	45.63	37.29	55.78
mBERT [22]	-	82.61	74.98	75.71	31.21	35.75	74.05	23.75	62.27	-
mBERT [9]	95.11	80.11	78.22	82.25	26.71	25.40	72.37	41.49	53.22	61.66
ZSJoint [29]	95.20	74.79	76.52	74.25	52.73	70.10	72.56	29.66	66.91	68.08
CoSDA [10]	92.29	81.37	76.94	79.36	64.06	66.62	75.05	48.77	77.32	73.47
GL-CLEF [7]	95.39	86.30	85.22	84.31	70.34	73.12	81.83	65.85	77.61	80.00
LAJ-MCL[8]	96.02	86.59	83.03	82.11	61.04	68.52	81.49	65.20	82.00	78.23
DiffSLU	**96.16**	**86.72**	**85.48**	**84.26**	**73.04**	**74.12**	**82.52**	**68.14**	**83.12**	**81.51**
Overall (Acc)	en	de	es	fr	hi	ja	pt	tr	zh	**AVG**
AR-S2S-PTR [31]	86.83	34.00	40.72	17.22	7.45	10.04	33.38	–	23.74	-
IT-S2S-PTR [32]	87.23	39.46	50.06	46.78	11.42	12.60	39.30	–	28.72	-
mBERT [9]	87.12	52.69	52.02	37.29	4.92	7.11	43.49	4.33	18.58	36.29
ZSJoint [29]	87.23	41.43	44.46	43.67	16.01	33.59	43.90	1.12	30.80	38.02
CoSDA [10]	77.04	57.06	46.62	50.06	26.20	28.89	48.77	15.24	46.36	44.03
GL-CLEF [7]	88.02	66.03	59.53	57.02	34.83	41.42	60.43	28.95	50.62	54.09
LAJ-MCL[8]	89.81	67.75	59.13	57.56	23.29	29.34	61.93	28.95	54.76	52.50
DiffSLU	**90.06**	**68.02**	**59.84**	**58.08**	**35.12**	**43.06**	**63.04**	**29.32**	**55.08**	**55.74**

Table 3: *Ablation results on the MultiATIS++ dataset.*

Models	Intent	Slot	Overall
DiffSLU	**92.90**	**81.51**	**55.74**
w/o diffusion	87.32(↓5.58)	73.47(↓8.04)	44.03(↓11.71)
More Parameters	88.24(↓4.66)	74.86(↓6.65)	45.12(↓10.62)
w/o distillation	92.53(↓0.37)	81.02(↓0.49)	55.16(↓0.58)

drops by 11.71%. These results demonstrate the importance of the diffusion module in our model, which achieves the mutual guidance between intent and slots.

3.5.2. Effect of More Parameters

Following previous works [2, 3], to verify whether the increased parameters of DiffSLU lead to the higher performance, we add an additional LSTM layer after the last layer of mBERT and refer it to *More Parameters*. The results in Table 3 show that our method outperforms mBERT with more parameters in intent accuracy, slot F1 and overall accuracy by 4.66%, 6.65%, 10.62%, respectively. These results demonstrate that the improvement of our method indeed comes from the diffusion module and knowledge distillation rather than the involved parameters.

3.5.3. Effect of Knowledge Distillation

To demonstrate the effectiveness of knowledge distillation, we remove it and refer it to *w/o CL* and the results are shown in Table 3. We can clearly observe that knowledge distillation is beneficial in improving the performance of the model, which plays a role in transferring knowledge from the model trained on the utterance in origin utterance to the model trained on the code-switched utterance. By applying knowledge distillation to facilitate the knowledge transfer between different languages, the model can predict the intent and slots more accurately.

4. Conclusions

In this paper, we propose a novel framework DiffSLU based on diffusion model and knowledge distillation for zero-shot cross-lingual spoken language understanding (SLU), which achieves the mutual guidance between intent and slots via the diffusion module. Besides, we apply knowledge distillation to further transfer knowledge. Experiments on MultiATIS++ dataset show that DiffSLU achieve a new state-of-the-art performance. Model analysis demonstrates that DiffSLU successfully transfers knowledge from source languages to target languages. In the future, we will explore the effectiveness of our method in other cross-lingual tasks to improve the performance.

5. Acknowledgements

The authors would like to express their sincere gratitude to the anonymous reviewers for their constructive feedback, which has helped to improve the quality of this paper significantly. This work was supported in part by the National Nature Science Foundation of China (grant numbers 71971067).

6. References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. of ICML*, 2015.

[12] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. of NeurIPS*, 2020.

[13] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. of ICLR*, 2021.

[14] X. Li, J. Thickstun, I. Gulrajani, P. S. Liang, and T. B. Hashimoto, "Diffusion-lm improves controllable text generation," in *Proc. of NeurIPS*, vol. 35, 2022, pp. 4328–4343.

[15] S. Gong, M. Li, J. Feng, Z. Wu, and L. Kong, "Diffuseq: Sequence to sequence text generation with diffusion models," *ArXiv preprint*, 2022.

[16] R. Strudel, C. Tallec, F. Altché, Y. Du, Y. Ganin, A. Mensch, W. Grathwohl, N. Savinov, S. Dieleman, L. Sifre *et al.*, "Self-conditioned embedding diffusion for text generation," *ArXiv preprint*, 2022.

[17] H. Yuan, Z. Yuan, C. Tan, F. Huang, and S. Huang, "Seqdiffuseq: Text diffusion with encoder-decoder transformers," *ArXiv preprint*, 2022.

[18] J. Lovelace, V. Kishore, C. Wan, E. Shekhtman, and K. Weinberger, "Latent diffusion for language generation," *ArXiv preprint*, 2022.

[19] S. Dieleman, L. Sartran, A. Roshannai, N. Savinov, Y. Ganin, P. H. Richemond, A. Doucet, R. Strudel, C. Dyer, C. Durkan *et al.*, "Continuous diffusion for categorical data," *ArXiv preprint*, 2022.

[20] Z. Gao, J. Guo, X. Tan, Y. Zhu, F. Zhang, J. Bian, and L. Xu, "Difformer: Empowering diffusion model on embedding space for text generation," *ArXiv preprint*, 2022.

[21] X. Han, S. Kumar, and Y. Tsvetkov, "Ssd-lm: Semi-autoregressive simplex-based diffusion language model for text generation and modular control," *ArXiv preprint*, 2022.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

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