HCLD: A Hierarchical Framework for Zero-shot Cross-lingual Dialogue System

Zhanyu Ma[1,2,3] **Jian Ye**[1,2,3*] **Xurui Yang**[1,2,3] **Jianfeng Liu**[1,2,3]
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China[1]
University of Chinese Academy of Sciences[2]
Beijing Key Laboratory of Mobile Computing and Pervasive Device[3]
{mazhanyu21s,jye,yangxurui20g,liujianfeng18s}@ict.ac.cn

Abstract

Recently, many task-oriented dialogue systems need to serve users in different languages. However, it is time-consuming to collect enough data of each language for training. Thus, zero-shot adaptation of cross-lingual task-oriented dialog systems has been studied. Most of existing methods consider the word-level alignments to conduct two main tasks for task-oriented dialogue system, i.e., intent detection and slot filling, and they rarely explore the dependency relations among these two tasks. In this paper, we propose a hierarchical framework to classify the pre-defined intents in the high-level and fulfill slot filling under the guidance of intent in the low-level. Particularly, we incorporate sentence-level alignment among different languages to enhance the performance of intent detection. The extensive experiments report that our proposed method achieves the SOTA performance on the public task-oriented dialog dataset.

1 Introduction

Natural language understanding (NLU) plays a significant role in task-oriented dialogue systems, which is aimed to parse dialog utterances by identifying user's intent and the arguments of the intent (Hou et al., 2021; van der Goot et al., 2021). These two tasks are known as intent detection and slot filling respectively. For instance, in the utterance "Next week's forecast" as shown in Table 1, the user intent is to query about the weather circumstances and the time argument of the query is "next week". Recently, many neural models have been proposed to jointly train these two tasks by considering the intent detection and the slot filling as sentence classification and sequence labeling task (Krishnan et al., 2021; Cao et al., 2020).

However, most works focus on monolingual datasets which are expensive to build. Furthermore, some dialogue systems, e.g., Google Home and Apple Siri, need to serve numerous users around the world in different languages and might be faced with the scarcity of dialogue data in certain languages. Thus, it's motivated to build cross-lingual dialogue systems that enable zero-shot adaptation from a high-resource language to a low-resource language without any training data in the target language. Specifically, some models (Upadhyay et al., 2018; Chen et al., 2018; Schuster et al., 2019) used cross-lingual pretrained embeddings to bridge different languages. Some approaches like Liu et al. (2019) adopted a small vocabulary of translated word-pairs to align cross-lingual embeddings by bilingual projection and alleviate the inherent discrepancies among different languages. Many existing methods (Liu et al., 2020; Qin et al., 2020) utilized pretrained cross-lingual language model (PXLM), e.g., XLM (Conneau and Lample, 2019), mBERT (Pires et al., 2019) and XLM-R (Conneau et al., 2020), to derive contextual embeddings of words in different languages.

Sentence	Next	week's	forecast
Slots	B-datetime	I-datetime	B-weather/noun
Intents	weather/find		

Table 1: An example of a sentence with slots and intents annotations from the zero-shot NLU dataset.

Nevertheless, most of existing methods consider embedding alignments among source and target languages at the word-level to fulfill the cross-lingual adaptation, but they ignore that intent detection is a sentence classification task (Liu and Lane, 2016; Liu et al., 2021), which requires higher level alignments among different languages. Furthermore, these two tasks are closely related and the slots highly depend on the intent (Goo et al., 2018; Qin et al., 2019), but their deep relations are not fully discovered by existing works.

In light of these observations, we propose a model named *HCLD* (A Hierarchical Framework

[*] Corresponding author.

Proceedings of the 29th International Conference on Computational Linguistics, pages 4492–4498
October 12–17, 2022.

for Zero-shot Cross-lingual Dialogue System) to alleviate these defects. Our approach is built upon a hierarchical framework to jointly accomplish intent detection and slot filling. It learns to classify pre-defined intents in the high-level of our model and fill the semantic slots under the guidance of the predicted intent in the low-level, which can help to find more related arguments of the intent to enhance the performance of slot filling. Particularly, we also adopt a pretrained language agnostic BiLSTM encoder *LASER* (Artetxe and Schwenk, 2019) to derive the sentence embeddings and directly use them for later classification task of intent detection, where the embeddings of sentences in different languages are aligned in the sentence-level. We conduct experiments on a public task-oriented dialogue dataset (Schuster et al., 2019). The results show that our proposed method achieves state-of-the-art performance on zero-shot adaptation.

2 Related Work

Zero-shot Cross-lingual Transfer The mainstream methods of zero-shot focus to learning cross-lingual embeddings. Recently, some contextual approaches (Pires et al., 2019; Conneau and Lample, 2019) built upon masked language model encourage to narrow the distance of representation in source and target language space. Apart from word-level alignments, another work (Artetxe and Schwenk, 2019) focuses on learning cross-lingual sentence embeddings to align sentences representations in different languages. MTOP (Li et al., 2021) further expend it with distant supervision in zero-shot setting for flat representations with the masked source utterance and the translated utterance as the concatenated input.

Code-switching is used as data augmentation in an alternative data alignment method known as CoSDA (Qin et al., 2020). To make model training highly multilingual, random words in the input are translated and substituted, leading to increased cross-lingual transfer ability. There were additional attempts to learn how to code-switch automatically (Liu et al., 2020).

Our work combines the cross-lingual sentence embedding and word embedding to handle intent detection and slot filling respectively.

Intent detection and slot filling Intent detection and slot filling are key tasks in the natural language understanding (NLU) of task-oriented dialogue systems. Recently, some models (Qin et al., 2021b; M'hamdi et al., 2021) consider to implement intent detection and slot filling jointly. The recent work (Zhang and Wang, 2016) first proposed a joint model to learn the correlation between intent and slots by RNN. Co-gat (Qin et al., 2021a) investigated a non-autoregressive model for joint multiple intent detection and slot filling.

Recently, information of intent has been incorporated for slot filling. The prior work (Goo et al., 2018) utilizes a slot-gated mechanism to model the relationship between two tasks. Some approaches like Wang et al. (2018) propose the bi-model to consider the cross-impact between the intent and slots based on ATIS datasets (Goo et al., 2018). Some models (Qin et al., 2019; Louvan and Magnini, 2020) predict the intent based on each word and then feed into slot filling as input.

Compared with previous works, our approach firstly predicts the intent given by multi-lingual sentence representation, and directly incorporates the intent information of sentence for slot filling. Furthermore, our model is handling simultaneously intent detection and slot filling at the cross-lingual setting.

3 Our Approach

3.1 Problem Formulation

Intent detection and slot filling are two key tasks of NLU. Given a utterance of L words $u = [w_1, w_2, \dots, w_L]$ and a set of pre-defined intent types I and slots types S, the intent detection is aimed to predict the intent $o^I \in I$ based on the utterance u. While the slot filling is a sequence tagging problem, mapping the word sequence $[w_1, w_2, \dots, w_L]$ into semantic slots $[s_1, s_2, \dots, s_L]$ where $s \in S$.

3.2 HCLD

As shown in Fig. 1, our proposed hierarchical model HCLD firstly classifies pre-defined intent in the high-level architecture and then fulfill the semantic slots under the guidance of intent in the low-level. For intent detection, multi-lingual sentence embedding feeds Sentence U into a sentence representation H_u. Then a linear layer would predict an intent i based on H_u. For slot filling, word sequence [1] of U, $u = w_1, w_2, \dots, w_L$ are passed

[1] https://huggingface.co/docs/transformers/v4.21.3/en/model_doc/xlm#transformers.XLMTokenizer

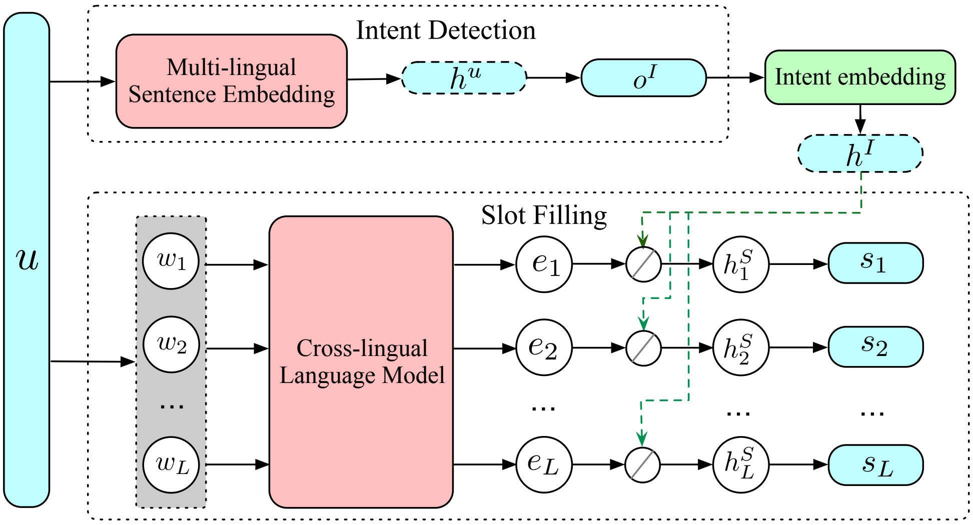


Figure 1: The framework of HCLD.

into multi-lingual language model [2] [3] and obtain contextual embeddings $e_1, e_2, \dots, e_{L-1}, e_L$. Next, H_s^j, $i \in \{1, \dots, L\}$ would computed by average the representations of e_j and H_i. Where H_i is obtained by intent i looking up the trainable intent embedding. Finally, the H_s^j would be used to predict the corresponding slot s_j.

In the high level, we first adopt LASER (Artetxe and Schwenk, 2019) [4] to derive sentence representations h^u. Based on multilingual neural machine translation tasks, LASER can produce sentence embeddings for multiple languages where embeddings of sentences with close meanings but in different languages can be aligned in a similar semantic space. Such sentence-level alignments can better boost the intent detection, which is a sentence classification task (Liu and Lane, 2016). Then the intent o^I is predicted by

$$y^I = Softmax(W_h^I h^u + b_h^I), \tag{1}$$

$$o^I = Argmax(y^I), \tag{2}$$

where y^I is the intent distribution and W_h^I, b_h^I are trainable parameters. In the low-level, we utilize mBERT (Pires et al., 2019), the pre-trained multilingual language model (PXLM), to produce contextual word embeddings in different languages $[e_1, e_2, \dots, e_L]$ where word embeddings are aligned in the same semantic space.

However, PXLM suffers inconsistent contextualized representations of subwords across different languages (Qin et al., 2020). Thus, we follow CoSDA-ML (Qin et al., 2020) [5] to better align subword representations with data augmentation during the fine-tuning process of mBERT.

Furthermore, we adopt the intent information to guide the slot filling task by averaging each word embedding e_j with the corresponding intent embedding h^I as new word representations $h_j^S = \text{Average}(e_j, h^I)$, and help to fill the semantic slot $[s_1, s_2, \dots, s_L]$. Here h_j^S is a combined representation with a word embedding and the intent embedding, and the slot would be predicted based on the h_j^S. Thus, the slot distribution for each word can be predicted as y_j^S. Recall that our proposed approach would guide the slot filling task with the information of the intent, we add a trainable intent embedding initialized with randomly parameters. A predicted intent o^I would look up the intent embedding and then receive an intent representation H^I. Next, H^I would influence the process of slot filling. The process of intent detection in the high-level can be formulated as follows,

$$H_u = LASER(U) \tag{3}$$

$$I_u = Softmax(Linear(H_u)) \tag{4}$$

Where I_u is a vector that tries to project the sentences in any language into a high-dimensional space with the goal that the same statement in any language will end up in the same neighborhood.

$$y_j^S = Softmax(W_h^S h_j^S + b_h^S), j \in \{1, \dots, L\} \tag{5}$$

To jointly learn both tasks, the objective function

[2] https://huggingface.co/bert-base-multilingual-cased

[3] https://github.com/facebookresearch/XLM/

[4] https://github.com/facebookresearch/LASER

[5] https://github.com/kodenii/CoSDA-ML

$\mathcal{L}$ is formulated as:

$$\mathcal{L} = \mathcal{L}_{\mathcal{I}} + \mathcal{L}_{\mathcal{S}} = -\sum_{i=1}^{n^I} \hat{y}_i^I log(y_i^I) - \sum_{j=1}^{L}\sum_{i=1}^{n^S} \hat{y}_{j,i}^S log(y_{j,i}^S) \tag{6}$$

where $\mathcal{L}_{\mathcal{I}}$, $\mathcal{L}_{\mathcal{S}}$ stand for the loss function of intent detection and slot filling respectively. n^I is the number of intent types and y_i^I is the gold intent label. While in $\mathcal{L}_{\mathcal{S}}$, L is the sequence length of a sentence, n^S means the number of slot types as well as y_j^S is the gold slot label. To avoid error prorogation, we adopt the gold intent to replace the predicted intent o^I during the training period.

3.3 Dataset

We conduct our experiments on the cross-lingual task-oriented dialogue dataset (Schuster et al., 2019) which contains English, Spanish and Thai. We train and validate our model on the English dialog data with $30,521$ and $4,181$ sentences respectively, and test on $3,043$ and $1,692$ sentences in Spanish and Thai for zero-shot adaptation following Liu et al. (2019).

3.4 Implementable Details

We introduce several competitive baselines, including: Zero-shot SLU (Upadhyay et al., 2018), BiLSTEM with CRF (Liu et al., 2019) [6] (LVM), XLM (Conneau and Lample, 2019), mBERT (Pires et al., 2019), Attention-Informed Mixed-Language Training (MLT) (Liu et al., 2020) [7] and CoSDA-ML (Qin et al., 2020). We adopt classification accuracy (Acc.) to evaluate the performance of intent detection while using F1 score to measure the performance of slot fillings.

In our experiments, we use WordPiece embeddings with a vocabulary containing 110k tokens following Devlin et al. (2019). We adopt LASER to generate sentence embeddings whose dimension is 1024 while taking the base case mBERT to derive word embeddings with dimension of 768.

Notice that we take the first subword embedding as word-level representation following Qin et al. (2020) while incorporating mBERT for slot filling. We also set the size of intent embedding to 768, then train our model for 10 epochs with a batch size of 32 and a learning rate 0.001. We adopt AdamW (Loshchilov and Hutter, 2018) to optimize our HCLD and select the hyper-parameters by grid search. Besides, we adopt the gold intent instead of the predicted intent o^I in equation 2 to guide the slot filling during the early stages of training period to avoid error prorogation.

4 Experiments

4.1 Overall Performance

We can make several observations from the results demonstrated in Table 2. Firstly, these methods with CoSDA-ML (*mBERT+COSDA-ML*) achieve the best performance among all the baselines, which indicates that pre-trained language models (e.g., mBERT) with data augmentation can produce better contextual word embeddings for different languages than those without data augmentation or only with simple word alignments techniques (e.g., BiLSTM with CRF/LVM).

Specifically, HCLD outperforms the second-best model on intent detection by 3.4 on Spanish, 1.5 on Thai. We conjecture that the alignments on sentence embeddings can enhance the adaption on different languages, leading to the significant improvements on intent detection. In addition, the performance of our model on slot filling also exceeds all the compared methods. We think that it can be attributed to the hierarchical architecture that fulfill the slots under the guidance of intent information.

4.2 Ablation Study

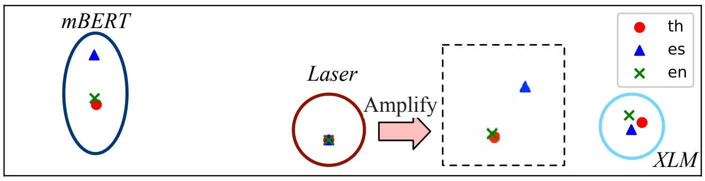
To investigate the effects of individual component, we conduct an ablation study and report the results in Table 2. Firstly, we remove the data augmentation method CoSDA-ML from our model to testify its effectiveness, it drops 11.2 and 16.3 on Spanish and Thai on slot filling respectively. Secondly, we remove the hierarchical architecture from HCLD and find that it would perform worse without the guidance from intent information. It is probably because intent can provide related information and help to find more accurate semantic slots. Thirdly, we investigate the importance of sentence alignments by replacing LASER with two pre-trained language models, i.e., mBERT or XLM, to derive the sentence embeddings. It validates the effectiveness of LASER that the performance of models with either mBERT or XLM is inferior to that with LASER which can produce aligned sentence representations in different languages. In addition, we also find that mBERT can produce better token-level embeddings for slot filling than XLM, thus

[6] https://github.com/zliucr/Crosslingual-NLU

[7] https://github.com/zliucr/mixed-language-training

Methods	Spanish		Thai	
	Intent Acc.	Slot F1	Intent Acc.	Slot F1
Zero-shot SLU (Upadhyay et al., 2018)	46.6	15.4	35.6	12.1
XLM (Conneau and Lample, 2019)	69.4	40.0	49.3	13.3
mBERT (Pires et al., 2019)	73.7	51.7	28.2	10.6
BiLSTM with CRF (Liu et al., 2019)	88.8	44.0	64.5	17.5
BiLSTM with LVM	89.2	64.0	70.8	29.5
XLM + MLT (Liu et al., 2020)	87.5	68.6	72.6	28.0
mBERT + MLT	87.9	74.9	73.5	27.1
XLM + CoSDA-ML (Qin et al., 2020)	90.3	69.0	*86.7*	34.9
mBERT +CoSDA-ML	*94.8*	*80.4*	76.8	*37.3*
HCLD	**98.2**	**83.2**	**88.2**	**38.0**
w/o CoSDA-ML	96.8	72.0	84.2	21.7
w/o Hierarchical	97.8	77.9	87.9	31.8
w/o LASER, w/ mBERT	84.7	79.5	81.0	32.2
w/o LASER, w/ XLM	93.0	69.4	86.3	34.4
w/o mBERT, w/ XLM	96.6	72.7	88.1	36.5

Table 2: Results of zero-shot test set are compared with different baselines as well as the result of ablation study. The most competitive results from baselines are annotated with underline.



th: ตอน เย็น นี้ อากาศ จะ เป็น อย่างไร

es: Cómo será el clima esta noche.

en: What will the weather be like this evening.

Figure 2: The visualization of sentence representations generated by LASER, XLM and mBERT. In the visulation, we select three utterances with the same meaning but in different languages, including Thai (th), Spanish (es) and English (en). In order to clearly demonstrate the points of LASER, we magnify these nodes with 10 times.

we derive word-embedding with mBERT in HCLD.

4.3 Visualization

To provide a more straightforward viewpoint to examine the sentence-level alignments, we visualize the sentence representations generated by LASER, mBERT and XLM by projecting them into a 2-dimension space. We select three sentences in English, Spanish and Thai with close meanings, which is an example from Liu et al. (2019). According to Fig. 2, we can find the three points from LASER are concentrated and even look like a point and while points from mBERT and XLM are more sparse. It validates that we can use LASER to derive aligned sentence embeddings and it has positive correlations with the results of intent detection as shown in Table 2.

5 Conclusion

In this paper, we propose a hierarchical framework to classify the pre-defined intents in the high-level and fulfill slot filling under the guidance of intent in the low-level. Particularly, we adopt sentence-level alignments to improve the performance of intent detection, and further enhance the performance of slot filling. The experiments conducted on the public dataset demonstrate the effectiveness of our proposed method and our model achieves state-of-the-art performance in the zero-shot cross-lingual scenario.

6 Acknowledgement

The research work is supported by National Key R&D Program of China (No.2017YFB1302400), Key Research and Development Program of in Shandong Province (2019JZZY020102), Industrial Internet Innovation and Development Project in 2021 (TC210A02M, TC210804D), Key Research and Development Program of Jiangsu Province (No.BE2018084), Opening Project of Beijing Key Laboratory of Mobile Computing and Pervasive Device.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Xu Cao, Deyi Xiong, Chongyang Shi, Chao Wang, Yao Meng, and Changjian Hu. 2020. Balanced joint adversarial training for robust intent detection and slot filling. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4926–4936.

Wenhu Chen, Jianshu Chen, Yu Su, Xin Wang, Dong Yu, Xifeng Yan, and William Yang Wang. 2018. Xl-nbt: A cross-lingual neural belief tracking framework. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 414–424.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757.

Yutai Hou, Yongkui Lai, Cheng Chen, Wanxiang Che, and Ting Liu. 2021. Learning to bridge metric spaces: Few-shot joint learning of intent detection and slot filling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3190–3200.

Jitin Krishnan, Antonios Anastasopoulos, Hemant Purohit, and Huzefa Rangwala. 2021. Multilingual code-switching for zero-shot cross-lingual intent prediction and slot filling. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 211–223.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962.

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

Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 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)*, pages 1297–1303.

Zihan Liu, Genta I Winata, Samuel Cahyawijaya, Andrea Madotto, Zhaojiang Lin, and Pascale Fung. 2021. On the importance of word order information in cross-lingual sequence labeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13461–13469.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8433–8440.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Samuel Louvan and Bernardo Magnini. 2020. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 480–496.

Meryem M'hamdi, Doo Soon Kim, Franck Dernoncourt, Trung Bui, Xiang Ren, and Jonathan May. 2021. X-metra-ada: Cross-lingual meta-transfer learning adaptation to natural language understanding and question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3617–3632.

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

Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu. 2019. A stack-propagation framework with token-level intent detection for spoken language understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference* *on Natural Language Processing (EMNLP-IJCNLP)*, pages 2078–2087.

Libo Qin, Zhouyang Li, Wanxiang Che, Minheng Ni, and Ting Liu. 2021a. Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13709–13717.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. *arXiv preprint arXiv:2006.06402*.

Libo Qin, Fuxuan Wei, Tianbao Xie, Xiao Xu, Wanxiang Che, and Ting Liu. 2021b. GL-GIN: Fast and accurate non-autoregressive model for joint multiple intent detection and slot filling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 178–188, Online. Association for Computational Linguistics.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038. IEEE.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021. From masked language modeling to translation: Non-english auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2479–2497.

Yu Wang, Yilin Shen, and Hongxia Jin. 2018. A bi-model based rnn semantic frame parsing model for intent detection and slot filling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 309–314.

Xiaodong Zhang and Houfeng Wang. 2016. A joint model of intent determination and slot filling for spoken language understanding. In *IJCAI*, volume 16, pages 2993–2999.