ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446325

A BIRGAT MODEL FOR MULTI-INTENT SPOKEN LANGUAGE UNDERSTANDING WITH HIERARCHICAL SEMANTIC FRAMES

Hongshen Xu[1*], *Ruisheng Cao*[1*], *Su Zhu*[2†], *Sheng Jiang*[2], *Hanchong Zhang*[1], *Lu Chen*[1], *Kai Yu*[1†]

[1]MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
[2]AISpeech Co., Ltd., Suzhou, China

ABSTRACT

Previous work on spoken language understanding (SLU) mainly focuses on single-intent settings, where each input utterance merely contains one user intent. This configuration significantly limits the surface form of user utterances and the capacity of output semantics. In this work, we firstly propose a Multi-Intent dataset which is collected from a realistic in-Vehicle dialogue System, called MIVS. The target semantic frame is organized in a 3-layer hierarchical structure to tackle the alignment and assignment problems in multi-intent cases. Accordingly, we devise a BiRGAT model to encode the hierarchy of ontology items, the backbone of which is a dual relational graph attention network. Coupled with the 3-way pointer-generator decoder, our method outperforms traditional sequence labeling and classification-based schemes by a large margin. Ablation study in transfer learning settings further uncovers the poor generalizability of current models in multi-intent cases.

Index Terms— Spoken Language Understanding, relational graph attention network, hierarchical semantic frame

1. INTRODUCTION

Spoken language understanding (SLU, [1]), which aims to parse the user utterance into a semantic frame, plays a critical role in building dialogue systems. Previous works focus on parsing utterances containing merely one intent. This simplification decomposes the original task into two sub-tasks, namely slot filling and intent detection [2]. When it comes to multi-intent cases [3], the traditional sequence labeling [4] (for slot filling) and sentence classification [5] (for intent detection) schemes are not applicable due to 1) the slot-value alignment problem, and 2) the slot-intent assignment issue.

Firstly, the same slot value may be aligned to multiple slots or used several times in the target semantic representation (duplicate alignments). As shown in Figure 1, the slot-value pair "`act=turn on`" is used twice to control the "`blue-tooth`" and "`music rhythm`". Besides, some frequently used slot values may be implicitly mentioned and will not occur as a continuous span in the input utterance, which is also known as the *unaligned slot value* problem [6]. In both cases, the traditional sequence labeling strategy is not applicable.

Furthermore, in multi-intent cases, if slot filling and intent detection are treated as separate tasks, the affiliation relationship from slot to intent can not be determined. In other words, slot-value pairs need to be clustered and allocated to their parent intent, restricted by the hierarchy of ontology items. As illustrated in Figure 1, we need to distinguish that the "`blue-tooth`" and "`front windows`" should be "`turned on` and "`closed`" respectively. And a simple modification of multi-class intent detection into multi-label classification [7, 8] will lose this slot-intent assignment information.

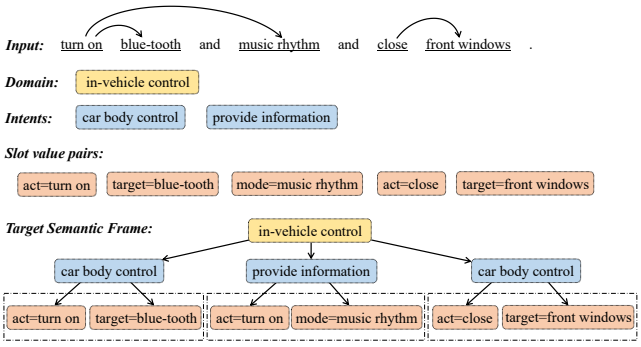


Fig. 1. A multi-intent example from MIVS dataset.

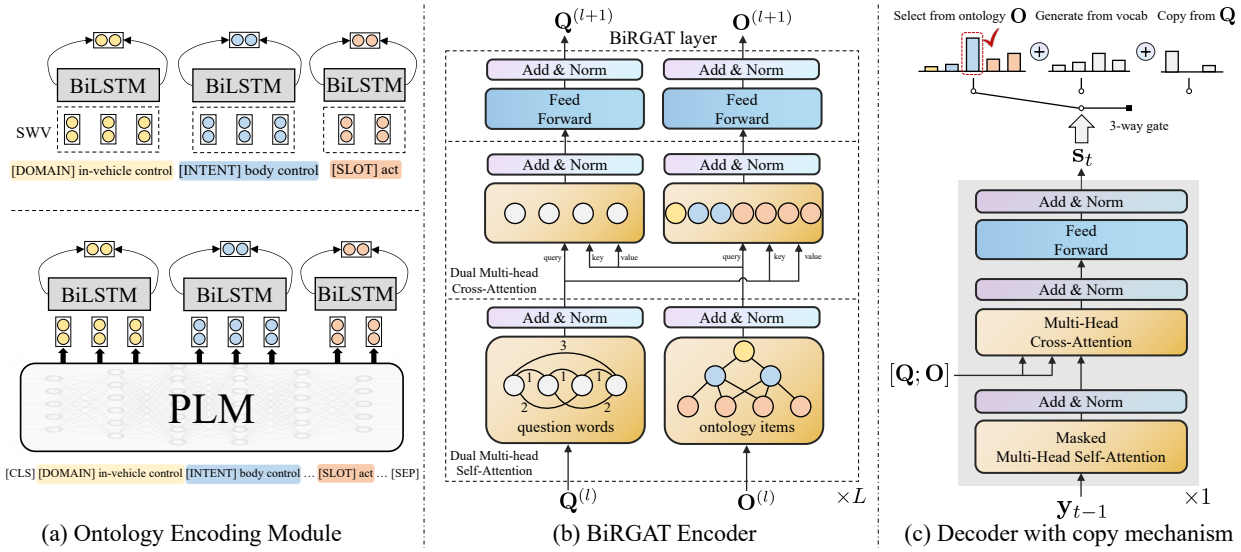
To this end, we firstly construct a large-scale Multi-Intent Chinese dataset collected from a realistic in-Vehicle System (MIVS) with $105,240$ data points. It also contains multi-domain samples where each input utterance involves two domains since users often lazily make their requests all at once for convenience. The target semantic frame is organized as a 3-layer tree, rooting from domains to intents and then slots (exemplified in the lower part of Figure 1). In accordance with this structured representation, we inject the hierarchy knowledge of ontology items into the encoder through two dual relational graph attention networks (RGAT, [9]). As for the decoder, after linearizing the tree representation into a string sequence with sentinal tokens, an adapted pointer-generator auto-regressive network [10] is utilized to selectively copy raw question words and ontology items to the output side. Experiments on two multi-intent datasets with hierarchical semantics, Chinese MIVS (this work) and English TOPv2 [11], demonstrate the advantage of our proposed BiRGAT framework over traditional methods. Codes and data are publicly available [1].

2. DATASET CONSTRUCTION

Given the ontology $O = \{o_i\}_{i=1}^{|O|}$ where o_i denotes a domain, intent, or slot, SLU converts an utterance $Q = (q_1, q_2, \cdots, q_{|Q|})$ into the semantic frame y. The hierarchical structure of domain→intent→slot is also provided as input structural priors.

* Equal Contribution.
† Su Zhu and Kai Yu are the corresponding authors.

[1]https://github.com/importpandas/MIVS_BIRGAT



(a) Ontology Encoding Module (b) BiRGAT Encoder (c) Decoder with copy mechanism

Fig. 2. An overview of the BiRGAT model architecture.

Prevalent benchmark ATIS [12] or SNIPS [13] simplifies the task by focusing on one single domain, considering one intent per utterance, and ignoring the hierarchy. In this work, we generalize to more practical scenarios where each utterance involves multiple intents and even multiple domains. Accordingly, the target semantic frame y is labeled as a tree to reflect the structure. The comparison to previous benchmarks is present in Table 1.

Dataset	Multi-domain	Multi-intent	Hierarchical Annotation	# of Samples
ATIS	✗	✗	✗	$6k$
SNIPS	✗	✗	✗	$14k$
MixATIS	✗	✓	✗	$20k$
MixSNIPS	✗	✓	✗	$50k$
TOPv2	✗	✓	✓	$181k$
MIVS (ours)	✓	✓	✓	$105k$

Table 1. Comparison to previous benchmarks.

The multi-intent MIVS dataset contains 5 different domains, namely *map*, *weather*, *phone*, *in-vehicle control* and *music*. The dataset can be split into two parts: **single-domain** examples contain both single-intent and multi-intent cases, which are collected and manually annotated from a realistic industrial in-vehicle environment; **cross-domain** examples are automatically synthesized following MixATIS [3]. Concretely, we extract two utterances from two different domains and concatenate them by conjunction words such as "and". The semantic tree is serialized as an output token sequence by inserting sentinel tokens such as brackets for clustering.

3. MODEL ARCHITECTURE

The entire BiRGAT model can be split into three parts as shown in Figure 2. Firstly, we adopt an ontology encoding module to obtain the initial ontology embedding (§ 3.1). Next, these features are further encoded via a dual RGAT for structural knowledge enhancement (§ 3.2). After that, an auto-regressive decoder is employed to produce the serialized semantic frame based on the encoded memory (§ 3.3).

3.1. Ontology Encoding Module

Inspired by the concept of label embeddings [14], given an ontology item $o_i^d \in O^d = \{o_i^d\}_{i=1}^{|O^d|}$ from one specific domain d (e.g., music), we can initialize its embedding $\mathbf{o}_i^d \in \mathbb{R}^{1\times m}$ (dimension is m) from either static word vectors (SWV) or pre-trained language models (PLMs) such as BERT [15], see Figure 2(a).

Concretely, we first concatenate all ontology items $o_i^d = (o_{i1}^d, \cdots, o_{il}^d)$ as well as their semantic type $o_{i0}^d \in \{\text{DOMAIN}, \text{INTENT}, \text{SLOT}\}$ to form a unified ontology sequence. Then the input sequence is fed into either a SWV module or PLM module to get token-level ontology embeddings. Finally, the forward and backward hidden states from a type-aware single-layer Bi-LSTM are concatenated as the ontology embedding $\mathbf{o}_i^d \in \mathbb{R}^{1\times m}$ for each ontology item o_i^d.

For domain d, we stack the initial embeddings of all ontology items to attain the domain feature matrix $\mathbf{O}^d \in \mathbb{R}^{|O^d|\times m}$. In multi-domain cases, we take one more step to stack matrices $\mathbf{O}^d$ from all possible domains d and get the entire matrix $\mathbf{O}^{(0)} \in \mathbb{R}^{|O|\times m}$, where $O = \bigcup_d O^d$. Otherwise, $\mathbf{O}^d$ directly serves as $\mathbf{O}^{(0)}$.

For input question $Q = (q_1, q_2, \cdots, q_{|Q|})$, the initial features $\mathbf{Q}^{(0)} \in \mathbb{R}^{|Q|\times m}$ can also be initialized from SWV or PLM.

3.2. BiRGAT Encoder

After obtaining the initial matrix $\mathbf{Q}^{(0)}$ and $\mathbf{O}^{(0)}$ of question words and ontology items, this module further enriches features with structural knowledge and cross-segment information. The BiRGAT encoder consists of L layers, the computation for layer l is

$$\mathbf{Q}^{(l+1)}, \mathbf{O}^{(l+1)} = \text{BiRGAT}(\mathbf{Q}^{(l)}, \mathbf{O}^{(l)}),$$

where $0 \le l \le L-1$. Each layer includes three sub-modules, namely 1) dual multi-head self-attention, 2) dual multi-head cross-attention, and 3) feedforward network. Each sub-module is also wrapped with residual connections and LayerNorm function.

3.2.1. Dual Multi-head Self-attention

Transformer [16] architecture is a specific implementation of graph attention network (GAT, [17]). To integrate the hierarchical structure among ontology items, we adapt the multi-head self-attention module by relative position embeddings [18]. Concretely, edge feature $\mathbf{z}_{ij}$ is introduced from adjacent ontology item o_j to o_i when computing the attention weight e_{ij} and attention vector $\tilde{\mathbf{o}}_i$,

$$e_{ij} = \frac{(\mathbf{o}_i\mathbf{W}_q)(\mathbf{o}_j\mathbf{W}_k + \mathbf{z}_{ij}\mathbf{W}_z)^{\mathrm{T}}}{\sqrt{m}},$$
$$\tilde{\mathbf{o}}_i = \sum_{j\in\mathcal{N}(i)} a_{ij}(\mathbf{o}_j\mathbf{W}_v + \mathbf{z}_{ij}\mathbf{W}_z),$$

where a_{ij} is the softmax version of e_{ij} and $\mathcal{N}(i)$ denotes the neighborhood of o_i.

We design the relation type z_{ij} between ontology items mainly to address the multi-intent and multi-domain problem. The slot items only have *slot-intent* relations with their parent intents. Similarly, the slot and intent items only have *slot-domain* and *intent-domain* relations with their parent domains, respectively. This design not only models the hierarchical semantic frame of the ontology structure but also mitigates the inter-domain information interference.

As for the question, we construct a complete graph among question words and utilize relative distances between words as the relation z_{ij}. To avoid over-parametrization, all edge features $\mathbf{z}_{ij}$ are shared across different layers and heads.

3.2.2. Dual Multi-head Cross-attention

This sub-module aims to gather information for each question word from the counterpart ontology items (and vice versa). By analogy to the multi-head cross-attention module in Transformer decoder, features of ontology items are incorporated as key/value vectors to enrich the representation of each question word. The symmetric part can be easily inferred, see the middle part of Figure 2(b).

After the relational graph encoding and cross-segment encoding, features of question words and ontology items are passed into a feedforward network. The outputs $\mathbf{Q}^{(L)}$ and $\mathbf{O}^{(L)}$ of the last layer L serve as the final encoder memory $\mathbf{Q}$ and $\mathbf{O}$.

3.3. Decoder with Copy Mechanism

Given encoder memory $\mathbf{Q}$ and $\mathbf{O}$, the output token sequence $y = (y_1, y_2, \cdots, y_{|y|})$ is produced auto-regressively via a single-layer Transformer decoder. The decoder hidden state $\mathbf{s}_t$ at timestep t is

$$\mathbf{s}_t = \text{TransformerDecoder}(\mathbf{y}_{<t}, [\mathbf{Q}; \mathbf{O}]).$$

Notice that y_t can be words in slot values, sentinel tokens (e.g., brackets), or an ontology item in the pre-defined specification. It is difficult to generate an ontology item token-by-token because a simple morphological change or synonym substitution will cause parsing errors while post-processing the linearized semantic frame y. Thus, we introduce a three-way gate to control the action of generating a token from a fixed vocabulary, copying a word from the question memory $\mathbf{Q}$, and selecting an ontology item from the ontology memory $\mathbf{O}$. Formally, given the decoder hidden state $\mathbf{s}_t \in \mathbb{R}^{1\times m}$,

$$\mathbf{g}_t = \text{softmax}(\mathbf{s}_t\mathbf{W}_g), \quad \mathbf{g}_t \in \mathbb{R}^{1\times 3},$$
$$P_{\text{gen}}(w_i) = \text{softmax}_i(\mathbf{s}_t\mathbf{W}_{\text{gen}}\phi(w_i)^{\mathrm{T}}),$$
$$P_{\text{copy}}(w_i) = \sum_{k:\ q_k=w_i} \text{PtrNet}(\mathbf{s}_t, \mathbf{Q})[k],$$
$$P_{\text{select}}(o_i) = \text{PtrNet}(\mathbf{s}_t, \mathbf{O})[i],$$
$$P(y_t|\mathbf{s}_t, \mathbf{Q}, \mathbf{O}) = g_{t1}P_{\text{gen}} + g_{t2}P_{\text{copy}} + g_{t3}P_{\text{select}},$$

where $\phi(w_i)$ returns the word embedding of w_i in a fixed vocabulary which is shared with the encoder, and $\text{PtrNet}(\mathbf{s}_t, \mathbf{Q})[k]$ denotes the probability of choosing the k-th entry (row) in memory $\mathbf{Q}$ which is implemented as the average weight from different heads of a multi-head cross-attention module (known as the pointer network, [10]). The training objective is decoupled as

$$\mathcal{L} = -\sum_{t=1}^{|y|} \log P(y_t|y_{<t}, \mathbf{Q}, \mathbf{O}).$$

4. EXPERIMENT

4.1. Datasets

We experiment on two multi-intent SLU datasets, namely Chinese MIVS (this work) and English TOPv2 [11]. The original output format of TOPv2 is inconsistent with our annotation. Thus, we convert the output labels of TOPv2 into our 3-layer hierarchical structure (§ 2). We report the sentence-level accuracy as the evaluation metric.

4.2. Implementation Details

Our model is implemented with Pytorch and `transformers` library. The hidden dimension m is $256/512$ for SWV and PLM respectively. The number of layers for the BiRGAT encoder is 2. As for the pointer-generator decoder, the number of layers is fixed to 1. The number of heads and dropout rate are set to 8 and 0.2 respectively. We use AdamW [19] optimizer with a linear warmup scheduler. The warmup ratio of total training steps is 0.1. The leaning rate and weight decay rate are $5e\text{-}4/1e\text{-}4$ for SWV and $2e\text{-}4/0.1$ for PLM. We train all the models with a batch size of 20 and $100k$ training iterations. For inference, we adopt beam search with size 5.

4.3. Main Results

In main experiments, we merge all data samples, including single-domain and multi-domain, and feed ontology items from all domains as input. Thus the model needs to determine the specific domain(s) of the current utterance. We adopt the classic Sequence Labeling (SL) method [20] as the baseline. By treating frequent but unaligned "(domain, intent, slot, value)" quadruples as utterance-level labels, we add a multi-label CLassiFier to tackle the unaligned slot value problem. From Table 2, we can observe that:

1) Compared to traditional methods SL and SL+CLF, sequence generation is more suitable for tackling the hierarchical semantic frames. The inherent limitation of SL-based methods makes it difficult to recover the semantic structure from a flattened sequence.

2) Ontology copy mechanism is significant to ensure the consistency with pre-defined names of ontology items. This conclusion is verified not only with our customized pointer-generator BiRGAT decoder but also with the end-to-end BART [21] model.

Init	Method	MIVS		TOPv2	
		Dev	Test	Dev	Test
SWV	SL	14.3	14.2	28.1	28.9
	+CLF	21.2	21.4	35.6	36.3
	BiRGAT	**84.9**	**85.6**	**86.1**	**85.9**
	w/o Copy	72.5	72.8	82.0	82.1
BART		27.0	26.8	84.3	83.7
w/ Copy		63.7	63.3	87.6	87.2
BERT	SL	14.8	14.9	29.1	29.8
	+CLF	23.0	23.1	36.8	37.5
	BiRGAT	**89.3**	**89.3**	**88.0**	**87.8**
	w/o Copy	75.6	75.5	84.1	84.1
RoBERTa	BiRGAT	89.3	89.2	88.1	87.9
ELECTRA		**90.0**	**90.2**	**88.4**	**88.0**

Table 2. Main Results on MVIS and TOPv2 datasets.

4.4. Ablation of BiRGAT Encoder

Init	w/ OE	GNN	w/ DCA	MIVS	TOPv2
SWV	✗	None	✗	83.3	82.6
	✓			83.9	85.3
	✓	GAT	✗	84.1	85.3
			✓	84.8	**85.9**
	✓	RGAT	✗	84.7	85.2
			✓	**85.6**	**85.9**
BERT	✗	None	✗	88.4	86.1
	✓			89.1	87.6
	✓	GAT	✗	88.4	87.7
			✓	**89.4**	**87.8**
	✓	RGAT	✗	89.1	87.3
			✓	89.3	**87.8**

Table 3. Ablation study on BiRGAT encoder. w/ OE: with ontology encoding; w/ DCA: with dual multi-head cross-attention.

In this section, we study the contribution of each component in the BiRGAT encoder, including Ontology Encoding module (OE), relational features $\mathbf{z}_{ij}$ (GAT v.s. RGAT), and Dual multi-head Cross-Attention sub-module of GNN layer (DCA). According to Table 3, we can summarize that:

1) Leveraging the text description of ontology items (w/ OE) is effective in enriching the semantics of ontology embeddings. This observation is consistent for both two datasets and all settings.

2) Both structural (GAT) and relational (RGAT) encoding can boost the performance. Compared to TOPv2, our MIVS dataset seems to benefit more from the hierarchy of ontology items. It can be attributed to the fact that data samples in MIVS contain relatively more intents and exhibit more complicated output tree structures.

3) Although the separate encoding of question and ontology already achieves remarkable results, the integration of cross-segment attention still brings stable performance gains on both datasets.

4.5. Transfer to More Intents

In this section, we explore the intent generalizability of current SLU models. Concretely, we train the model on data samples with the number of intents less or equal to 3, but evaluate the model on examples containing more intents (> 3). We can further fine-tune the model with a few samples containing > 3 intents. We conduct experiments on domain "*in-vehicle control*" since it contains more intents on average. We also introduce a large language model (LLM, [22]) baseline, i.e., `text-davinci-003` with in-context learning (ICL) for comparison. From Figure 3 we can observe that:

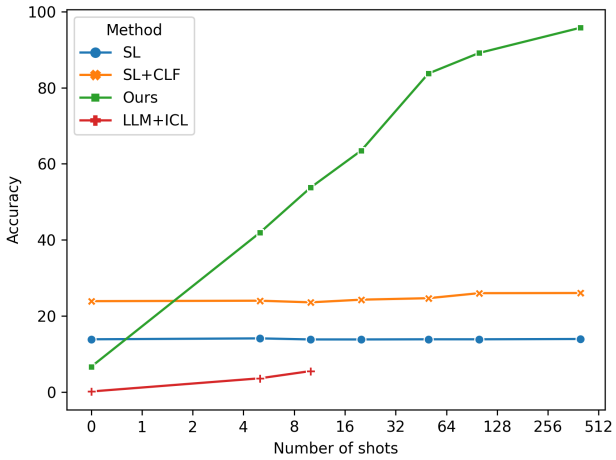


Fig. 3. Few-shot learning experiments when transferring to more intents (> 3) in domain "*in-vehicle control*". Due to the max token limit, prompts of LLM are truncated to at most 10 exemplars.

1) Disappointingly, our method is less performant than SL-based methods in zero-shot settings. Through the case study, we find that most erroneous predictions merely contain 3 intents and omit an entire intent sub-tree. We hypothesize that the generation-based method suffers from the problem of over-fitting the output length, while SL-based methods achieve better generalization of length variation.

2) In few-shot settings, our method dramatically surpasses SL-based methods with merely 5 samples. Moreover, SL-based methods attain limited improvements from fine-tuning. It can be explained that outputs with more intents usually present more complicated structures, which may be tough for SL-based methods to reconstruct.

3) Despite exciting results in other fields, it is difficult for LLM to produce both semantically coherent and syntactically-valid output sequences with very few exemplars.

5. CONCLUSION

In this work, we propose a large-scale multi-domain multi-intent dataset MIVS which is collected from a realistic in-vehicle dialogue system. Accordingly, we devise a BiRGAT model to incorporate the hierarchy of ontology items into the graph encoder and introduce a three-way copy mechanism to the decoder. Experiments on datasets MIVS and TOPv2 demonstrate the superiority of BiRGAT over various baselines.

6. ACKNOWLEDGEMENTS

This work is funded by the China NSFC Projects (92370206, U23B2057, 62106142 and 62120106006) and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).

7. REFERENCES

[1] Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams, "Pomdp-based statistical spoken dialog systems: A review," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1160–1179, 2013.

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

[3] Libo Qin, Xiao Xu, Wanxiang Che, and Ting Liu, "AGIF: An adaptive graph-interactive framework for joint multiple intent detection and slot filling," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online, Nov. 2020, pp. 1807–1816, Association for Computational Linguistics.

[4] Su Zhu and Kai Yu, "Encoder-decoder with focus-mechanism for sequence labelling based spoken language understanding," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5675–5679.

[5] Rashmi Gangadharaiah and Balakrishnan Narayanaswamy, "Joint multiple intent detection and slot labeling for goal-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)*, Minneapolis, Minnesota, June 2019, pp. 564–569, Association for Computational Linguistics.

[6] Zijian Zhao, Su Zhu, and Kai Yu, "A hierarchical decoding model for spoken language understanding from unaligned data," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*. 2019, pp. 7305–7309, IEEE.

[7] Bowen Xing and Ivor Tsang, "Co-guiding net: Achieving mutual guidances between multiple intent detection and slot filling via heterogeneous semantics-label graphs," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 159–169.

[8] Bowen Xing and Ivor W Tsang, "Group is better than individual: Exploiting label topologies and label relations for joint multiple intent detection and slot filling," *arXiv preprint arXiv:2210.10369*, 2022.

[9] Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang, "Relational graph attention network for aspect-based sentiment analysis," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020, pp. 3229–3238, Association for Computational Linguistics.

[10] Abigail See, Peter J. Liu, and Christopher D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, July 2017, pp. 1073–1083, Association for Computational Linguistics.

[11] Xilun Chen, Asish Ghoshal, Yashar Mehdad, Luke Zettlemoyer, and Sonal Gupta, "Low-resource domain adaptation for compositional task-oriented semantic parsing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, Association for Computational Linguistics.

[12] Charles T Hemphill, John J Godfrey, George R Doddington, et al., "The atis spoken language systems pilot corpus," in *Proc. the DARPA speech and natural language workshop*, 1990, pp. 96–101.

[13] Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau, "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces," *CoRR*, vol. abs/1805.10190, 2018.

[14] Su Zhu, Zijian Zhao, Rao Ma, and Kai Yu, "Prior knowledge driven label embedding for slot filling in natural language understanding," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1440–1451, 2020.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019, pp. 4171–4186, Association for Computational Linguistics.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[17] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[18] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani, "Self-attention with relative position representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana, June 2018, pp. 464–468, Association for Computational Linguistics.

[19] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. 2019, OpenReview.net.

[20] Libo Qin, Tailu Liu, Wanxiang Che, Bingbing Kang, Sendong Zhao, and Ting Liu, "A co-interactive transformer for joint slot filling and intent detection," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 8193–8197.

[21] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880.

[22] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe, "Training language models to follow instructions with human feedback," *CoRR*, vol. abs/2203.02155, 2022.