



Modeling Multi-Turn Spoken Language Understanding with Dynamic Graph Convolutional Networks

Yi Huang, Si Chen, Jingyu Yao, Junlan Feng*

JIUTIAN Team, China Mobile Research Institute, China

{huangyi, chensiyjy, yaojingyu, fengjunlan}@chinamobile.com

Abstract

Spoken Language Understanding (SLU) stands as a pivotal element within task-oriented dialogue systems, where it leverages the dialogue context to steer intent detection at the utterance level and slot filling at the token level. Nonetheless, the challenge of judiciously assimilating dialogue context into multi-turn SLU persists as a formidable hurdle. This difficulty is compounded by the inherently dynamic distribution of conversational information in real-world settings, where key details, such as shifts in intent and behavior, can be overshadowed by a deluge of less critical data. To address this issue, we introduce a novel SLU model tailored for intent detection and slot filling in multi-turn interactions. Central to our approach is the incorporation of a dynamic graph convolutional network that selectively amalgamates essential historical information into the dialogue context, thereby enhancing the model's sensitivity to contextually relevant cues. We subject our model to rigorous evaluation on three widely recognized benchmark datasets: SIM, CMCC and FewJoint. The experimental outcomes underscore the model's superior performance, and further results from real-scene datasets strengthen the effectiveness of the method.

Index Terms: spoken language understanding, task-oriented dialogue systems, graph convolutional network

1. Introduction

Task-oriented dialogue represents a pivotal research domain, garnering significant interest from both academic and industrial sectors. Within this domain, SLU serves as a crucial interface between users and machines. The primary objective of SLU is to facilitate the completion of routine tasks for users, such as booking seats or tickets [1]. Key tasks of SLU include intent detection [2, 3] and slot filling [4, 5], which operate at the utterance and token levels, respectively. The intent detection task is viewed as an utterance-level classification task, while the slot filling task can be formulated as a token-level sequence labeling task [6, 7]. Both components are instrumental in predicting user objectives and identifying essential slot values throughout the dialogue's progression. Due to the close correlation between intent detection and slot filling, predominant approaches [8, 9, 10] in the literature involve employing joint models to consider the relationship between the two tasks.

Compared to single-turn SLU, multi-turn SLU [11, 12, 13, 14] boasts a broader spectrum of applications. However, it also encounters a significant challenge: the local information crucial for understanding is often overshadowed by an overwhelming quantity of other information. Most models [15, 16] rely on complex heuristic aggregation functions to integrate context

information, which necessitates specialized knowledge and restricts the model's flexibility and generalization capability.

To address this issue, we introduce a novel SLU model. This model employs a dynamic graph convolutional network, eliminating the need for heuristic aggregation functions to amalgamate contextual information. Given the dynamic nature of information distribution, our model selectively incorporates key historical data into the context. This enables a fine-grained integration of contextual information to guide utterance-level intent detection and token-level slot filling.

Our contributions are twofold: (1) We propose a dynamic graph convolutional network that adaptively refines node features based on the dialogue's progression, allowing for a more nuanced understanding of the conversation's state. To enhance our model, we further incorporate a PMI-based loss function to address issues brought by the imbalance nature of dialogue information. (2) We conduct extensive experiments to prove the advantages of proposed framework on the benchmark datasets and compare them with existing methods. Experimental results show our framework achieves state-of-the-art performance.

2. Methodology

In this section, we describe two main components of our proposed model: an input-encoder that encodes both utterance-level and token-level information simultaneously, and a dynamical graph convolutional network that uses independent encoding to dynamically merge and detect important information in the joint tasks.

2.1. Input Encoding and Pooling

First, all dialogues are passed to the input encoding and pooling step for processing. We construct a graph structure in the following way.

Nodes: Each discourse representation in a dialogue is set to a node. For $u \in [1, 2, \dots, T]$, each node is initialized with the corresponding sequentially encoded feature vector e_u . Thus, the first level state vector of all nodes is $E^0 = (e_1^0, \dots, e_T^0)$, where T denotes the number of discourses in the dialogue.

Edges: We construct a dialogue graph in which each node is interconnected so that contextual features can be propagated from neighbouring nodes to the current node.

2.2. Dynamic Graph Convolution

Our graph convolution network takes the initial representation of subsection 2.1 as input and outputs an aggregated dialogue representation through a dynamic updating network, as shown in Figure 1.

*Corresponding author.

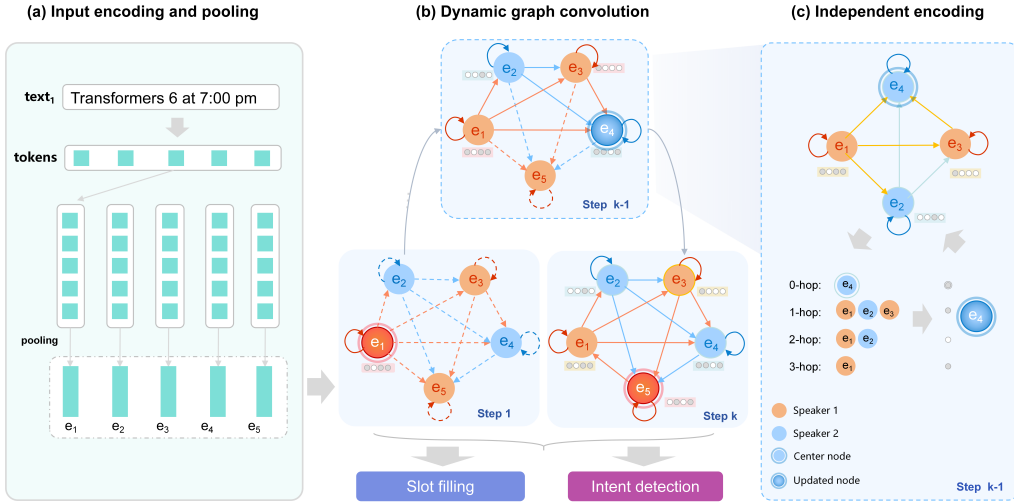


Figure 1: The illustration of our proposed framework. Part (a) tokenizes and pools the input multiple contextual utterances to obtain a set of sentence sequences; Part (b) introduces the process of dynamically updating node representations, where the current node incorporates only nodes from the historical and current utterances as neighboring nodes for graph construction; Part (c) provides a detailed explanation of how the representation of the central node (marked with double circles in the graph) is influenced by its adjacent nodes after graph construction.

2.2.1. Independent Encoding

To get the final node representation of the dialogue graph, this paper uses an independent encoding step on the central node and its neighborhoods from different hops. We encode potentially important nodes using independent encoding and stitch the node with the encoding of its k -hop neighbours ($1 \leq k \leq K$) to represent each node in the graph. When performing graph convolution operations, each node representation is updated with its historical neighbours. Therefore information about the relationships between nodes can be captured while avoiding the injection of uncertain information in the future. In the form of:

$$n_u = n_u^{(0)} \oplus n_u^{(1)} \oplus \dots \oplus n_u^{(k)}, \quad (1)$$

where n_u is the property representation of this node, and $n_u^{(k)}$ is the representation composed of the k -hop neighbourhood nodes of node u . Specifically, $n_u^{(0)}$ and $n_u^{(k)}$ are represented as:

$$n_u^{(0)} = \kappa(e_u^{(0)}), \quad (2)$$

$$n_u^{(k)} = \kappa(A_u E_u^{(k-1)} W^{(k)})_u, \quad (3)$$

where A_u is the adjacency matrix of the graph network, $E_u^{(k-1)}$ represents the representation composed of k -hop neighbourhood nodes, κ is the activation function, and $W^{(k)}$ is the trainable weight matrix.

And the entire graph representation can then be obtained as follows:

$$h_{G_i} = R(n_u | u \in V_i), \quad (4)$$

where $R(\cdot)$ is a sum-based or mean-based global pooling operation. V_i is the set of all nodes. The following tasks of the intent detection and the slot filling can then be carried out in conjunction with the contextual information obtained earlier.

2.2.2. Loss function

Our loss function is divided into three components: intent detection loss, slot filling loss and the loss with PMI (Pointwise Mutual Information) [17], which are updated by joint optimization.

The intention detection loss and the slot filling loss take the following forms respectively:

$$\mathcal{L}_I \triangleq - \sum_{i=1}^m \hat{y}^{(i,I)} \log(y^{(i,I)}), \quad (5)$$

$$\mathcal{L}_S \triangleq - \sum_{i=1}^m \sum_{j=1}^{n_j} \hat{y}_j^{(i,S)} \log(y_j^{(i,S)}), \quad (6)$$

where m is the number of training data, n_j is the number of tokens in the j th data, $\hat{y}^{j,I}$ is the target intention label, and $\hat{y}_j^{(i,S)}$ is the target slot label.

In the traditional graph encoding, certain node representations that play a crucial role but constitute a minority may be overshadowed by representations from the majority of other nodes. To address this issue, we model the PMI between input graphs and their node representations. Specifically, we use $f(G, \theta)$ to denote the proposed SLU model, where θ represents the model parameters. And $P_\theta(G_i, y_i)$ is the distribution of y_i given G_i . Modeling the PMI between the input graph G_i and its important/unimportant information can be expressed as:

$$f_{y_i}(G_i, \Theta) \sim \log \frac{P_\Theta(G_i, y_i)}{P(G_i)P(y_i)}, \quad (7)$$

where the right part of the formula represents the PMI between G_i and y_i , and $P(y_i)$ is the proportion of graphs in the training set T with the label y_i . In contrast to many existing loss functions, the PMI enables our model to prioritize learning crucial information that reveals the intrinsic correlations between

graphs and their labels. Equation 7 can be reformulated as follows:

$$\log P_{\Theta}(y_i|G_i) \sim f_{y_i}(G_i, \theta) + \log P(y_i). \quad (8)$$

Introducing P_{θ} into our loss function, we can obtain the PMI-based loss function:

$$\begin{aligned} \mathcal{L}_M \triangleq & -\frac{1}{m} \sum_{i=1}^m (\log P_{\theta}(y^{(i)}|G_i) + \\ & \frac{1}{n_j} \sum_{j=1}^{n_j} \log P_{\theta}(y_j^{(i)}|N_j)), \end{aligned} \quad (9)$$

where N_j is the j th input node of graph G_i .

The final loss function is:

$$\mathcal{L} = \alpha_1 \mathcal{L}_I + \alpha_2 \mathcal{L}_S + \alpha_3 \mathcal{L}_M, \quad (10)$$

where α_1, α_2 and α_3 are hyper-parameters.

3. Experiment

In this section, we present experiments to provide empirical validation of our proposed SLU model. We compare our model with existing baseline methods, using diverse evaluation metrics, including intent accuracy and slot F1.

3.1. Dataset and Baselines

The dataset Situation model and Information-seeking Management, SIM [18], includes two subtypes (Sim-R and Sim-M). This dataset contains task-based English dialogues about different scenarios such as travel reservations, restaurant reservations, etc. And we further explore the task performance on two Chinese datasets in real-life scenarios, CMCC [19] from the user-customer service conversations and FewJoint [20] from the AIUI open dialogue platform of iFlytek¹.

We conduct a comparison of our model against several challenging baselines:

- NoContext: A two-layer bidirectional RNN that uses GRU and LSTM cells without any contextual information.
- EfficientNet [21]: A hierarchical recurrent neural network that efficiently encodes dialogue act context.
- GraphNet [22]: A graph convolutional network for integrating dialogue act context.
- ContextFusion [23]: A contextual SLU model with an adaptive global-local context fusion mechanism for multi-turn intent detection and slot filling.
- GPT-4o [24]: An advanced conversational AI model utilizing the GPT (Generative Pre-trained Transformer) architecture to generate human-like text based on the input it receives.
- Baichuan4 [25]: A LLM (Large Language Model)-based product boasting over one hundred billion parameters, demonstrating superior performance in several medical evaluations and general ability tests.
- SereTOD [26]: A task-oriented dialogue system construction incorporated with information extraction task from dialogue transcripts.

¹<http://aiui.xfyun.cn/index-aiui>

Table 1: Main results on the English dataset (SIM) with different baselines and our framework, representing the average performance across 10 independent experiments conducted under identical conditions. The DynGCN in the last line means the dynamic graph convolution network.

Model	Sim-R (%)		Sim-M (%)	
	Intent Acc	Slot F1	Intent Acc	Slot F1
NoContext	83.61	94.24	88.51	86.91
EfficientNet	99.65	94.70	99.27	93.73
GraphNet	99.97	95.37	99.93	94.41
ContextFusion	99.97	98.10	100	96.98
SereTOD	98.84	94.40	100	92.11
GPT-4o	98.84	87.09	98.96	89.84
Baichuan4	99.74	85.23	99.24	87.01
Our model	99.97	98.36	100	97.18
w/o DynGCN	97.76	94.97	98.83	90.58

3.2. Implementation Details

For Independent Encoding in subsection 2.1, we initialize BERT models with Huggingface Models Hub checkpoints^{2,3} and fine-tune it with the provided configuration using the joint training scheme presented in the paper on the SIM, CMCC and FewJoint. We train the model on a single machine with one Tesla V100 GPU for 28 hours and choose the checkpoint with the lowest cross-entropy loss term on a held-out validation set. The hyperparameter hop K is identically set to 3 to all the experiments. We study the effects of various numbers of K in Equation 1 and do not find our approach sensitive to its choice and omit it from the equation.

For LLM-based models, we employ the function calling method to perform the intent detection and slot filling tasks on the SIM, CMCC and FewJoint. Based on the slot descriptions and the function calling construction documentation from the official OpenAI documentation⁴, we construct the functions of slot filling and intent detection for each dataset used in the experiment.

3.3. Result and Analysis

Sim-R and Sim-M: The main evaluation results of the simulated English dialogue dataset are shown in Table 1. This demonstrates that our proposed dynamic convolution network is effective for multi-turn SLU, especially in the task of slot filling. For the task of intent detection, LLMs reach nearly an accuracy of 99% on the dataset SIM. This is because LLMs possess powerful language understanding capabilities and strong generalization abilities. The simulated datasets are constructed by algorithms rather than extracted from real conversations, so the semantics are relatively simple, which leads to decent results on the dataset SIM without fine-tuning. In the slot filling task, our proposed model achieves an F1 score of 98.36% on the Sim-R dataset and 97.18% on the Sim-M dataset. However, existing LLMs typically achieve F1 scores ranging from 85% to 87% on Sim-R and from 87% to 90% on Sim-M. These results suggest that our model significantly outperforms prior work in the slot

²<https://huggingface.co/bert-base-uncased>

³<https://huggingface.co/bert-base-chinese>

⁴<https://platform.openai.com/docs/guides/function-calling>

Table 2: Main results on the Chinese datasets (CMCC and FewJoint) with different baselines and our framework, representing the average performance across 10 independent experiments conducted under identical conditions.

Model	CMCC (%)		FewJoint (%)	
	Intent Acc	Slot F1	Intent Acc	Slot F1
SereTOD	68.60	34.95	61.79	62.96
GPT-4o	54.42	54.66	75.87	77.08
Baichuan4	48.14	54.37	76.28	76.79
Our model	72.15	57.09	80.06	80.73
w/o DynGCN	59.70	36.89	48.86	59.75

filling task. The overall results on the English simulated dataset demonstrate that our proposed dynamic convolution network is effective in the evaluation metrics of intent detection and slot filling for multi-turn SLU.

CMCC and FewJoint: Given that the English dataset (SIM) is derived from simulated scenarios, the complexity of the intent detection task is relatively low. Therefore, we further explore the general performance in real dialogue scenarios. We conduct experiments on two Chinese dialogue datasets, CMCC and FewJoint. On the CMCC dataset, our model achieves an accuracy of 72.15% on the intent detection task. This performance outperforms the former baselines and the representative LLMs, which typically attain accuracies ranging from 48% to 54%. The performance of intent detection task indicates a strong ability to discern the underlying intent of customer queries. Additionally, the F1 score for slot filling reaches the highest 57.09%, showcasing the model’s proficiency in extracting key information such as packages and messages in the field of mobile communication. On another dataset from real-world scenarios, FewJoint, our model achieves a performance advantage of over 3% in both tasks of the intent detection and the slot filling. The FewJoint dataset has few-shot samples for each domain, which is suited for few-shot joint SLU tasks, reflects the real-world complexity beyond slot-filling and intent detection. Since the datasets from real-scene scenarios contain more complex semantic and label information, the result determines that our proposed dynamic convolution network is effective in the evaluation metrics of multi-turn conversation.

3.4. Ablation Study

To investigate the impact of the dynamical network on the overall performance of our model, an ablation study is conducted by replacing the dynamic network with a two-layer multi-layer perceptron. The results in the last line of Table 1 and 2 reveal a significant degradation in performance. This observation underscores the importance of the dynamical convolutional network in enhancing the model’s predictive capabilities and highlights its critical role in achieving optimal performance.

Furthermore, we conduct experiments varying with the quantity of pre-given samples on the dataset FewJoint, as shown in Figure 2. As training samples decreases, the performance of our model maintains a lead of 1–9% over the LLM. Under data-scarce conditions, the performance loss of our model (27%) is significantly lower than the LLM (34%), indicating that our model has a lower dependency on the training set. As the size of the training data diminishes, the slot filling task experiences a more substantial impact (39%) in contrast to the intent detection task (29%).

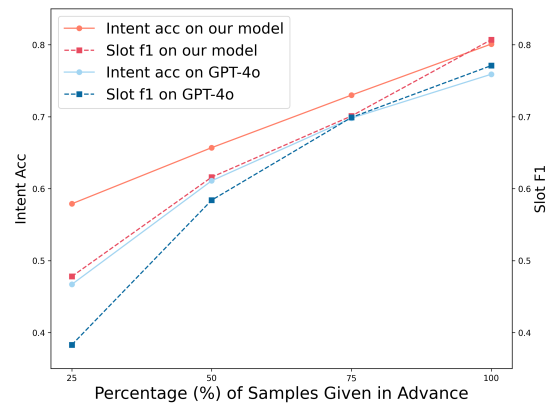


Figure 2: Ablation results on the performance variation with the quantity of pre-given samples on the dataset FewJoint.

To analyse the comparative impact of multi-task joint learning and independent training of each task, we conduct corresponding experiments on the English dataset. As shown in Figure 3, each bar represents a unique experimental condition, delineating either a multi-task or independent learning approach applied to one of the datasets. The vertical axis of the bar chart measures the performance scores, which are expressed in terms of accuracy and F1 score. The visual representation clearly demonstrates that the multi-task learning approach consistently outperforms the independent training approach across both intent detection and slot filling tasks, as evidenced by the higher scores attributed to the multi-task condition.

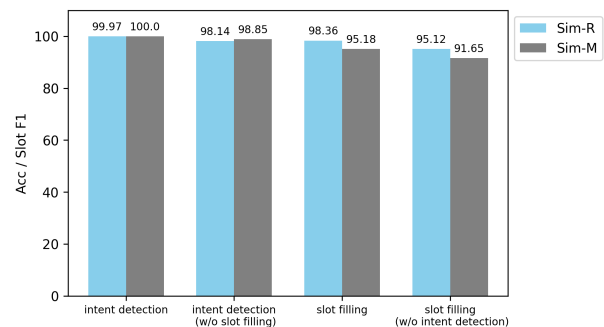


Figure 3: Ablation results on Sim-R and Sim-M datasets. From left to right, the first and third one are under the joint training setting, whereas the other two are not.

4. Discussion

This paper presents a focused and effective model for intent detection and slot filling tasks in multi-turn dialogue scenarios. The model can well adhere to the natural flow of human conversation and address the problem of dynamic nature of dialogue graph in spoken language understanding tasks by using a dynamic graph convolution network. After experimental validation, our model achieves good results on the SLU tasks surpassing previous work, especially for the datasets from real-world scenarios. For future work, we plan to explore techniques that adaptively infuse potential dialogue structural information into the graph network through LLMs.

5. Acknowledgements

This work was supported by the Beijing Natural Science Foundation (L222006) and China Mobile Holistic Artificial Intelligence Major Project Funding (R22105ZS, R22105ZSC01).

6. References

- [1] S. Louvan and B. Magnini, "Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey," *arXiv preprint arXiv:2011.00564*, 2020.
- [2] C. Xia, C. Zhang, X. Yan, Y. Chang, and P. S. Yu, "Zero-shot user intent detection via capsule neural networks," 2018.
- [3] J. Liu, Y. Li, and M. Lin, "Review of intent detection methods in the human-machine dialogue system," in *Journal of physics: conference series*, vol. 1267, no. 1. IOP Publishing, 2019, p. 012059.
- [4] H. Shi, "A sequence-to-sequence approach for numerical slot-filling dialog systems," in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2020, pp. 272–277.
- [5] J. Liu, M. Yu, Y. Chen, and J. Xu, "Cross-domain slot filling as machine reading comprehension: A new perspective," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 673–685, 2022.
- [6] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang, "Disan: Directional self-attention network for rnn/cnn-free language understanding," 2017.
- [7] Z. Tan, M. Wang, J. Xie, Y. Chen, and X. Shi, "Deep semantic role labeling with self-attention," 2017.
- [8] X. Zhang and H. Wang, "A joint model of intent determination and slot filling for spoken language understanding," 2016.
- [9] L. Qin, W. Che, Y. Li, H. Wen, and T. Liu, "A stack-propagation framework with token-level intent detection for spoken language understanding," 2019.
- [10] Y. Liu, F. Meng, J. Zhang, J. Zhou, Y. Chen, and J. Xu, "Cmnet: A novel collaborative memory network for spoken language understanding," 2019.
- [11] H. Bai, Y. Zhou, J. Zhang, and C. Zong, "Memory consolidation for contextual spoken language understanding with dialogue logistic inference," *arXiv preprint arXiv:1906.01788*, 2019.
- [12] K. Wei, T. Tran, F.-J. Chang, K. M. Sathyendra, T. Muniyappa, J. Liu, A. Raju, R. McGowan, N. Susanj, A. Rastrow *et al.*, "Attentive contextual carryover for multi-turn end-to-end spoken language understanding," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 837–844.
- [13] X. Cheng, W. Xu, Z. Zhu, H. Li, and Y. Zou, "Towards spoken language understanding via multi-level multi-grained contrastive learning," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023, pp. 326–336.
- [14] L. Cheng, W. Yang, and W. Jia, "A scope sensitive and result attentive model for multi-intent spoken language understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 12 691–12 699.
- [15] W. Peng, Y. Hu, L. Xing, Y. Xie, Y. Sun, and Y. Li, "Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation," *ArXiv*, vol. abs/2204.12749, 2022.
- [16] Q. Chen, Z. Zhuo, and W. Wang, "Bert for joint intent classification and slot filling," *ArXiv*, vol. abs/1902.10909, 2019.
- [17] L. Ren, M. Sidhu, Q. Zeng, R. G. Reddy, H. Ji, and C. Zhai, "Cpmi: conditional pointwise mutual information for turn-level dialogue evaluation," *arXiv preprint arXiv:2306.15245*, 2023.
- [18] P. Shah, D. Z. Hakkani-Tür, G. Tür, A. Rastogi, A. Bapna, N. N. Kennard, and L. Heck, "Building a conversational agent overnight with dialogue self-play," *ArXiv*, vol. abs/1801.04871, 2018.
- [19] Y. Huang, X. Wu, S. Chen, W. Hu, Q. Zhu, J. Feng, C. Deng, Z. Ou, and J. Zhao, "Cmcc: a comprehensive and large-scale human-human dataset for dialogue systems," in *Proceedings of the Towards Semi-Supervised and Reinforced Task-Oriented Dialog Systems (SereTOD)*, 2022, pp. 48–61.
- [20] Y. Hou, J. Mao, Y. Lai, C. Chen, W. Che, Z. Chen, and T. Liu, "Fewjoint: A few-shot learning benchmark for joint language understanding," *arXiv preprint arXiv:2009.08138*, 2020.
- [21] R. Gupta, A. Rastogi, and D. Hakkani-Tur, "An efficient approach to encoding context for spoken language understanding," *arXiv preprint arXiv:1807.00267*, 2018.
- [22] L. Qin, W. Che, M. Ni, Y. Li, and T. Liu, "Knowing where to leverage: Context-aware graph convolutional network with an adaptive fusion layer for contextual spoken language understanding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1280–1289, 2021.
- [23] T. Tran, K. Wei, W. Ruan, R. McGowan, N. Susanj, and G. P. Strimel, "Adaptive global-local context fusion for multi-turn spoken language understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, 2022, pp. 12 622–12 628.
- [24] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.-L. Han, and Y. Tang, "A brief overview of chatgpt: The history, status quo and potential future development," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 5, pp. 1122–1136, 2023.
- [25] A. Yang, B. Xiao, B. Wang, B. Zhang, C. Bian, C. Yin, C. Lv, D. Pan, D. Wang, D. Yan *et al.*, "Baichuan 2: Open large-scale language models," *arXiv preprint arXiv:2309.10305*, 2023.
- [26] H. Liu, H. Peng, Z. Ou, J. Li, Y. Huang, and J. Feng, "Information extraction and human-robot dialogue towards real-life tasks: A baseline study with the mobilecs dataset," *arXiv preprint arXiv:2209.13464*, 2022.