A UNIFIED FRAMEWORK FOR MULTI-INTENT SPOKEN LANGUAGE UNDERSTANDING WITH PROMPTING

Feifan Song1,2 Lianzhe Huang3 Houfeng Wang1,2

1 School of Computer Science, Peking University
2 National Key Laboratory of Multimedia Information Processing, Peking University
3 Pattern Recognition Center, WeChat AI, Tencent

ABSTRACT

ChatGPT has demonstrated impressive capabilities in building conversations. However, for Spoken Language Understanding (SLU) with multiple intents, traditional approaches where Intent Detection and Slot Filling are jointly modeled with distinct formulations hinder networks from effectively extracting shared features. In this work, we describe a Prompt-based SLU (PromptSLU) framework, to intuitively unify two sub-tasks into the same form for a common pre-trained model. Specifically, variable intents are predicted first, then naturally embedded into prompts to guide slot-value inference from a semantic perspective. Furthermore, we are inspired by multi-task learning to introduce an auxiliary sub-task and a concise general objective, which helps to learn relationships among provided labels. Experiment results show that our framework outperforms several competitive baselines on two datasets. The source code is available at https://github.com/F2-Song/PromptSLU.

Index Terms— Spoken Language Understanding, Multi-intent Classification, Slot Filling, Prompting

1. INTRODUCTION

The great success acquired by ChatGPT released from OpenAI has inspired the community to involve general large language models (LLMs) in open-domain dialogue. However, diverse modules in task-oriented dialogue (TOD) remain to be further considered. For example, Spoken Language Understanding (SLU) serves as an entrance to the TOD implementation pipeline. It typically consists of two sub-tasks: Intent Detection (ID) and Slot Filling (SF). The former is usually modeled into a text classification problem and the latter is completed by cutting out fragments from an input utterance in sequence tagging.

Early works [1, 2, 3, 4] may focus on one sub-task of the two or treat them separately, but the correlation between them has been the key to further improvement of SLU in recent years [5, 6, 7, 8, 9]. Moreover, real-world utterances can be longer and contain multiple intents. In Figure 1, the utterance “Show me the cheapest fare in the database and ground transportation san francisco” contains two distinct intents (atis_cheapest and atis_ground_service). For multi-intent SLU like this, similar joint modeling methods for ID and SF are discussed [10, 11, 12, 13, 14, 15], and potential interaction among intents and slots plays an essential role.

Although following the path of joint modeling, prior methods tend to integrate ID and SF in a typical NLU way, by providing a shared feature extractor during encoding but treating them individually when decoding. Figure 1 displays different decoding processes of two sub-tasks as multi-label classification and sequence tagging.

![Figure 1](https://github.com/F2-Song/PromptSLU)

The vast distinction in task formulation between ID and SF is indispensable. It discourages models from effective shared feature extraction and in turn, hinders the potential of comprehensively better performance. Consequently, unifying the two sub-tasks from start to finish with a common framework is significant.

In this paper, we adopt another paradigm different from previous NLU works by proposing an NLG-based framework, called Prompt-based Spoken Language Understanding (PromptSLU) framework. It leverages general pre-trained language models (PLMs) and prompting, rather than delicate network structures, to construct the correlation among labels across ID and SF, which handles the aforementioned challenges. Given an utterance and task requirement, PromptSLU fills the utterance in a task-specific prompt template, inputs it into the PLM, and generates corresponding labels.

Besides, consistency between two sub-tasks is crucial in SLU. The multi-intent scenario is challenging for accurate alignment from intents to slots due to the greater length of utterances and increased number of labels. For this issue, we follow the sequential way that intents are predicted first to guide SF. We transform it into an intuitive formulation in NLG: plugging intents into prompts to restrain SF from a semantic perspective, namely Semantic Intents Guidance (SIG). As is plain to humans in form, this design also allows an explicit interaction between ID and SF, outside of the common PLM. Furthermore, inspired by multi-task learning used in [16, 17], we propose an auxiliary sub-task, called Slot Prediction (SP), to steer models to additionally maintain semantic consistency. We also provide a concise general objective called Split Loss, which is applied to ID, SF and SP, in order to better improve unified formulation and heuristically maintain balance among diverse sub-tasks. Experiments demonstrate that PromptSLU outperforms competitive baselines on most metrics, including those using PLMs. Abundant analyses also exploit interesting features of PromptSLU.
2. APPROACH

PromptSLU utilizes prompts to complete different sub-tasks, which is an intuitive way similar to [16], but is specially designed for multi-intent SLU containing a particular Semantic Intent Guidance for intent-slot interaction. We also propose a new auxiliary sub-task Slot Prediction to improve Intent Detection (ID) and Slot Filling (SF) and a concise Split Loss to heuristically balance different objectives. The pipeline and human-crafted prompts are shown in Figure 2.

2.1. Initial Modeling

Given the utterance $X = \{x_1, x_2, x_3, ..., x_N\}$, for each of sub-tasks, PromptSLU inputs to the sharing pre-trained language model (PLM) the concatenation of a task-specific prefix and $X$, then makes predictions in a common form, which is crucial for the purpose of integrating task formulations. Different from [8, 17], we use natural language (NL) to represent non-semantic intent labels, reducing the difficulty of exploiting semantic information.

As to ID, intents are traditionally defined as a fixed number of labels. A model is fed with input text $X$, then maps it to these labels. Differently, we model the task in the form of text generation, i.e., the PLM produces a sequence of intents $I_X$ corresponding to a human-crafted prefix.

For SF, raw data have to be processed in advance to let the PLM generate slot-value pairs directly. For each sample, we first extract slot-value pairs according to the original BIO-tagged sequence, based on tagging rules. These golden pairs are then orderly assembled into one sentence, which serves as a text-generation target $SF_X$. Similar to that in ID, we also replace slot labels with human-readable NL phrases.

2.2. Semantic Intent Guidance

It is noted that there exists semantic similarity among tokens, intents and slots. This is reflected in the aspects of not only frequent co-occurrence but also semantic resemblance, which is helpful for inference. We consider it in our framework.

We adopt joint modeling by parsing intents from the output of ID and using them to facilitate other sub-tasks, which we name as the Semantic Intent Guidance (SIG) mechanism and display in (4) of Figure 2. Together with task-specific prefixes, intents also serve as an essential part of prompts for other sub-tasks, to guide their completion and keep semantic consistency among basic elements.

2.3. Slot Prediction

Multi-task learning is apt to offer a channel for the interaction of different tasks. Inspired by [16, 17], we design this auxiliary sub-task in the training stage, forcing the PLM to focus on the semantic interaction among tokens, intents, and slots. We leverage it for the improvement of maintaining semantic consistency, which is also enhanced after inserting the SIG.

2.4. Training

Similar to other text-to-text tasks, our framework is trained to minimize negative log-likelihood for each sub-task:

$$-\sum_{i=1}^{[Y]} \log p_{\Theta}(y_i|y_{<i}, X)$$

where $Y$ denotes the target and $\Theta$ denotes model parameters. We introduce two variants of loss functions.
Table 1. Main results. Methods are distinguished by whether pre-trained language models are involved.

| Methods          | MixATIS          | MixSNIPS         |
|------------------|------------------|------------------|
|                  | Slot(F1) | Intent(Acc) | Overall(Acc) | Slot(F1) | Intent(Acc) | Overall(Acc) |
| Non-pre-trained  |           |             |             |          |             |             |
| SF-ID [9]        | 87.4     | 66.2        | 34.9        | 90.6     | 95.0        | 59.9        |
| Stack-Propagation [6] | 87.8     | 72.1        | 40.1        | 94.2     | 96.0        | 72.9        |
| Joint Multiple ID-SF [10] | 84.6     | 73.4        | 36.1        | 90.6     | 95.1        | 62.9        |
| AGIF [11]        | 86.7     | 74.4        | 40.8        | 94.2     | 95.1        | 74.2        |
| GL-GIN [13]      | 88.3     | 76.3        | 43.5        | 94.9     | 95.6        | 75.4        |
| SDJN [14]        | 88.2     | 77.1        | 44.6        | 94.4     | 96.5        | 75.7        |
| Pre-trained      |           |             |             |          |             |             |
| SDJN+BERT [18]   | 87.5     | 78.0        | 46.3        | 95.4     | 96.7        | 79.3        |
| GL-GIN+RoBERTa [19] | 88.6     | 79.2        | 53.9        | 95.9     | 97.4        | 82.4        |
| GL-GIN+T5\text{large} | 88.8     | 79.2        | 52.4        | 96.2     | 96.9        | 83.4        |
| SLIM [15]        | 88.5     | 78.3        | 47.6        | 96.5     | 97.2        | 84.0        |
| PromptSLU\text{\text{large}}$_\text{base}$ | 88.6 | 83.6 | 53.3 | 96.2 | 97.7 | 82.8 |
| PromptSLU\text{\text{large}}$_\text{full}$ | 89.6 | 85.8 | 57.2 | 96.5 | 97.5 | 84.8 |

Table 2. Ablation experiments. SP and SIG denote Slot Prediction and Semantic Intent Guidance, respectively.

| Settings          | Methods          | Slot(F1) | Intent(Acc) | Overall(Acc) |
|-------------------|------------------|----------|-------------|--------------|
|                   | PromptSLU\text{\text{large}}$_\text{full}$ | 87.1     | 80.4        | 50.6        |
|                   | PromptSLU\text{\text{large}}$_\text{base}$ | 89.6     | 85.8        | 57.2        |
| Original dataset  | SP               | 87.9     | 82.6        | 50.8        |
|                   | SIG              | 88.8     | 84.6        | 53.3        |
|                   | only ID/SF       | 88.6     | 81.4        | N/A         |
| + golden ID       | PromptSLU\text{\text{large}}$_\text{full}$ | 87.7     | 80.4        | 50.6        |
|                   | PromptSLU\text{\text{large}}$_\text{base}$ | 89.7     | 85.8        | 57.2        |

3. EXPERIMENT

3.1. Settings

We conduct experiments mainly on MixATIS and MixSNIPS [11], which are widely used multi-intent SLU datasets. As clarified previously, we extract intents and slot-value pairs for each sample. We evaluate the performance of Slot Filling with F1 score, Intent Detection and sentence-level semantic frame parsing with accuracy, named Slot F1, Intent Acc, and Overall Acc. The baselines involved can be found in Table 1.

In this work, we choose T5 [20] as the backbone of PromptSLU with small (60M parameters) and base (220M parameters) versions, because the prompt style is coherent with the pre-trained tasks of T5.

3.2. Main Results

Table 1 displays the experimental results of our framework and baselines. PromptSLU outperforms each baseline on each metric for both MixATIS and MixSNIPS, no matter whether a pre-trained model is leveraged. In general, pre-trained models are beneficial to this task. It mainly comes from the utilization of knowledge contained. Among them, PromptSLU is distinguished from other NLU methods by more concise modeling but better performance. This result verifies the efficient knowledge exploitation of the proposed prompt-based text generation paradigm for multi-intent SLU. Specifically, PromptSLU\text{\text{large}}$_\text{base}$ accomplishes significant advances over the results of baselines, especially those from GL-GIN+T5\text{large} with a more powerful backbone.

3.3. Ablation Study

We abstract several factors and analyze their effectiveness and results are shown in Table 2. Due to limited space, we only place results on MixATIS in this part, which proves the most intuitive idea that model size is a dominant factor. Besides, after removing Slot Prediction (SP) or the SIG mechanism, there are consequent drops on every metric, especially on Overall Acc. It confirms that SP and SIG serve as influential catalysts to help maintain semantic consistency between two kinds of labels, which is also one of the chief strengths of jointly modeling. Actually, if ID and SF are separately completed, performance decreases similarly.

We surprisingly find that the score of Intent Acc also decreases without SIG. We attribute it to some implicit semantic supervision during the update of model parameters, i.e., although parameters gradient cannot be propagated across sub-tasks, our framework still has the potential to revise intents prediction against SIG.

Furthermore, transmission error should be considered in SIG. We replace predicted intents with golden ones and see that PromptSLU\text{\text{large}}$_\text{base}$ obviously beats PromptSLU\text{\text{large}}$_\text{full}$ on almost all metrics, except close value of Intent Acc for MixSNIPS. It primarily presents the superiority of \( \mathcal{L}_s \) against the straightforward \( \mathcal{L}_w \) on balancing different objectives. For other experiments, we use \( \mathcal{L}_s \) as the loss by default.
### Table 3. Results on MixATIS and MixSNIPS with different low-resource settings. We randomly sample 1%, 5% and 30% training data, and compare PromptSLU with two representative baselines GL-GIN and GL-GIN+RoBERTa.<br><br>![PromptSLU results](image)

#### 3.4. Low-resource Study

Along with the incredible few-shot or zero-shot performance of GPT-3 [21], the potential of PLMs is widely recognized in low-resource scenarios, due to knowledge acquired in the pre-training stage. However, how to implement PLMs for efficient knowledge exploitation is crucial. In this part, we conduct experiments to compare PromptSLU with GL-GIN and its promoted version, as shown in Table 3. We obtain three observations from the results:

i) The performance of the three methods rises along with the expanding scale of training data. Among them, PromptSLU\textsubscript{base} and GL-GIN+RoBERTa\textsubscript{base} generally surpass the non-pre-trained GL-GIN: It displays the effect of increasing data volume and PLMs on steadily boosting performance.

ii) Surprisingly, GL-GIN+RoBERTa\textsubscript{base} does not work better as expected than GL-GIN on SF under the low-resource setting, though achieves a higher Intent Acc. It suggests the pre-trained RoBERTa\textsubscript{base} indeed helps improve the quality of utterance embedding, which is key to ID; but fails to produce better token-level representations that are responsible for SF.

iii) PromptSLU\textsubscript{base} surpasses both baselines on MixATIS of each scale. Despite the nuance of Intent Acc on 1% and 30% MixSNIPS between PromptSLU\textsubscript{base} and GL-GIN+RoBERTa\textsubscript{base}, PromptSLU\textsubscript{base} excels with higher Slot F1 and Overall Acc. These prove PromptSLU gets efficient access to intrinsic knowledge in PLMs, which compensates for the loss of data scale.

#### 3.5. Effectiveness of Semantics

NL descriptions are utilized to better understand the semantics of labels. To explore the effectiveness of this operation, we replace them with corresponding raw labels (as special tokens). Figure 3 illustrates the comparison between the two strategies. We set the same hyper-parameters as those for the main results and make the following observations:

i) Despite using PLMs, Figure 3 shows a great drop in Overall Acc without NL descriptions. The absence of semantics hinders PromptSLU from aligning intents with slots.

ii) Figure 3 further demonstrates its dynamic influence on SF and ID. Comparing the convergence rates of two strategies on Slot F1 and Intent Acc, we see that PS\textsubscript{n} quickly reaches a relatively high level. However, PS\textsubscript{r} evolves step by step at early stages, along with the modification of label representations subject to the local context. It suggests that semantics efficiently help fine-tuning.

In this paper, we propose PromptSLU, which handles Intent Detection and Slot Filling in multi-intent SLU with a prompt-based text generation framework. To the best of our knowledge, this is the first work to utilize prompts for this problem. Our framework integrates these two sub-tasks into the same formulation while distinguishing them according to diverse prompts, which simplifies the architecture of whole framework. Moreover, based on the semantic similarity between intents and slots, predicted intents are driven to guide the process of Slot Filling, while a new auxiliary task Slot Prediction is introduced. Experimental results show its multi-dimensional superiority against baselines on two datasets.

### 4. CONCLUSION

This work was supported by the National Natural Science Foundation of China (Grant No. 62036001), and the National Key R&D Program of China (No. 2022ZD0116308).
6. REFERENCES

[1] Suman Ravuri and Andreas Stolcke, “Recurrent neural network and lstm models for lexical utterance classification,” in Sixth Annual Conference of the International Speech Communication Association, 2015.

[2] Byeonchang Kim, Seonghan Ryu, and Gary Geunbae Lee, “Two-stage multi-intent detection for spoken language understanding,” Multimedia Tools Appl., vol. 76, no. 9, pp. 11377–11390, may 2017.

[3] Gabor Fuisz, Ivan Vulić, Samuel Gibbons, Inigo Casanueva, and Paweł Budzianowski, “Improved and efficient conversational slot labeling through question answering,” arXiv preprint arXiv:2204.02123, 2022.

[4] Shikib Mehri and Maxine Eskenazi, “Gensf: Simultaneous adaptation of generative pre-trained models and slot filling,” arXiv preprint arXiv:2106.07055, 2021.

[5] Linhao Zhang and Houfeng Wang, “Using bidirectional transformer-crf for spoken language understanding.” in Natural Language Processing and Chinese Computing, Jie Tang, Min-Yen Kan, Dongyan Zhao, Sujian Li, and Hongying Zan, Eds., Cham, 2019, pp. 130–141, Springer International Publishing.

[6] Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu, “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), Hong Kong, China, Nov. 2019, pp. 2078–2087, Association for Computational Linguistics.

[7] Linhao Zhang, Yu Shi, Linjun Shou, Ming Gong, Houfeng Wang, and Michael Zeng, “A joint and domain-adaptive approach to spoken language understanding,” arXiv preprint arXiv:2107.11768, 2021.

[8] Panagiotis Tassias, “A prompting-based encoder-decoder approach to intent recognition and slot filling,” M.S. thesis, Athens University of Economics and Business, 2021.

[9] Haihong E, Peiqing Niu, Zhongfu Chen, and Meina Song, “A novel bi-directional interrelated model for joint intent detection and slot filling,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, July 2019, pp. 5467–5471, Association for Computational Linguistics.

[10] 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.

[11] 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.

[12] Zeyuan Ding, Zhihao Yang, Hongfei Lin, and Jian Wang, “Focus on interaction: A novel dynamic graph model for joint multiple intent detection and slot filling,” in Proceedings of the 20th International Joint Conference on Artificial Intelligence, 2021, pp. 3801–3807.

[13] Libo Qin, Fuxuan Wei, Tianbao Xie, Xiao Xu, Wanxiang Che, and Ting Liu, “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), Online, Aug. 2021, pp. 178–188, Association for Computational Linguistics.

[14] Lisong Chen, Peilin Zhou, and Yuexian Zou, “Joint multiple intent detection and slot filling via self-distillation,” in 2022 IEEE International Conference on Acoustics, Speech and Signal Processing, 2022, ICASSP ’22, pp. 7612–7616.

[15] Fengyu Cai, Wanhao Zhou, Fei Mi, and Boi Faltings, “Slim: Explicit slot-intent mapping with bert for joint multi-intent detection and slot filling,” in 2022 IEEE International Conference on Acoustics, Speech and Signal Processing, 2022, ICASSP ’22, pp. 7607–7611.

[16] Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang, “Multi-task pre-training for plug-and-play task-oriented dialogue system,” in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, May 2022, pp. 4661–4676, Association for Computational Linguistics.

[17] Liwen Wang, Rumei Li, Yang Yan, Yuennmeng Yan, Sirui Wang, Wei Wu, and Weiran Xu, “Instructionner: A multi-task instruction-based generative framework for few-shot ner,” arXiv preprint arXiv:2203.03903, 2022.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[19] Yinhao Liu, Myle Ott, Naman Goyal, Jingfei Yan, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” arXiv preprint arXiv:1907.11692, 2019.

[20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” Journal of Machine Learning Research, vol. 21, no. 140, pp. 1–67, 2020.

[21] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei, “Language models are few-shot learners,” in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 1877–1901, Curran Associates, Inc.

9970