Zero-Shot Spoken Language Understanding via Large Language Models: A Preliminary Study

Zhihong Zhu[†], **Xuxin Cheng**[†], **Hao An**,
Zhichang Wang, **Dongsheng Chen**, **Zhiqi Huang**[*]
School of Electronic and Computer Engineering, Peking University
{zhihongzhu, chengxx, anhao, wzcc, chends}@stu.pku.edu.cn,
zhiqihuang@pku.edu.cn

Abstract

Zero-shot Spoken Language Understanding (SLU) aims to enable task-oriented dialogue systems to understand user needs without training data. Challenging but worthwhile, zero-shot SLU reduces the time and effort that data labeling takes. Recent advancements in large language models (LLMs), such as GPT3.5 and ChatGPT, have shown promising results in zero-shot settings, which motivates us to explore prompt-based methods. In this study, we investigate whether strong SLU models can be constructed by directly prompting LLMs. Specifically, we propose a simple yet effective two-stage framework dubbed GPT-SLU, which transforms the SLU task into a question-answering problem. Powered by multi-stage mutual guided prompts, GPT-SLU can leverage the correlations between two subtasks in SLU to achieve better predictions, which is greatly explored in the traditional fine-tuning paradigm. Experimental results on three SLU benchmark datasets demonstrate the significant potential of LLMs for zero-shot SLU. Comprehensive analyses validate the effectiveness of our proposed framework and also indicate that there is still room for further improvement of LLMs in SLU scenarios.

Keywords: Task-oriented Dialogue System, Spoken Language Understanding, Large Language Model

1. Introduction

Spoken Language Understanding (SLU) constitutes a pivotal component in task-oriented dialogue systems, which aims to extract semantic information from user utterances (Qin et al., 2021). Recent advancements in SLU have led to successful applications across various industries, including voice assistants and voice-controlled smart devices (Chen et al., 2022). To be specific, SLU comprises two subtasks: intent detection, which identifies users' intents, and slot filling, which extracts semantic constituents from the user's query. Considering the high correlations between these two subtasks, joint training models have been proposed (Xing and Tsang, 2022; Zhu et al., 2024, 2023b) and have shown promising results. However, mainstream SLU models heavily rely on supervised training using labeled data. Working with an enormous amount of labeling data is invariably hectic, labor-intensive, and time-consuming. Consequently, numerous attempts have focused on fine-tuning techniques to minimize manual labor with zero/few-shot methods, *e.g.*, few-shot SLU (Wu et al., 2021; Hou et al., 2022) and zero-shot cross-lingual SLU (Qin et al., 2022; Zhu et al., 2023a; Cheng et al., 2023b).

Recently, the advancement of Large Language Models (LLMs), such as GPT-3 (Brown et al., 2020), InstructGPT (Ouyang et al., 2022) and ChatGPT, has significantly accelerated progress in the field of Natural Language Processing (NLP) (Chen et al., 2024). Among them, ChatGPT excels in various NLP tasks, such as summarization (Yang et al., 2023), machine translation (Jiao et al., 2023b) and information extraction (Wei et al., 2023). Therefore, a timing question arises: *Is it also effective to prompt LLMs to do zero-shot SLU tasks?*

More recently, Pan et al. (2023); He and Garner (2023); Li et al. (2023) conducted an initial evaluation of the potential of ChatGPT for SLU. However, the correlations between the two subtasks have been overlooked when utilizing LLMs to address SLU, leading to suboptimal performance. Our core insight is to *exploit the correlations between the two subtasks to address SLU under the LLM-based framework, similar to the fine-tuning paradigm*.

In this paper, we explore the capabilities of ChatGPT and hypothesize that it inherently possesses qualities suitable for developing a zero-shot SLU model interactively. Concretely, we present a simple yet effective two-stage framework GPT-SLU, which transforms the SLU task into a question-answering problem. In the first stage, GPT-SLU aims to generate the initial intent and slot sequence for the input utterance. Then in the second stage, GPT-SLU utilizes intent and slots from stage one as cues to mutually guide each other. By doing this, GPT-SLU enables two subtasks to guide each other, akin to traditional fine-tuned joint models, and to some extent alleviates the hallucination issue (Wang et al., 2023; Huang et al., 2024) in LLMs.

We conduct experiments on three widely used SLU benchmarks including ATIS (Hemphill et al., 1990), SNIPS (Coucke et al., 2018) and

[†]Equal Contribution.
[*]Corresponding author.

LREC-COLING 2024, pages 17877–17883

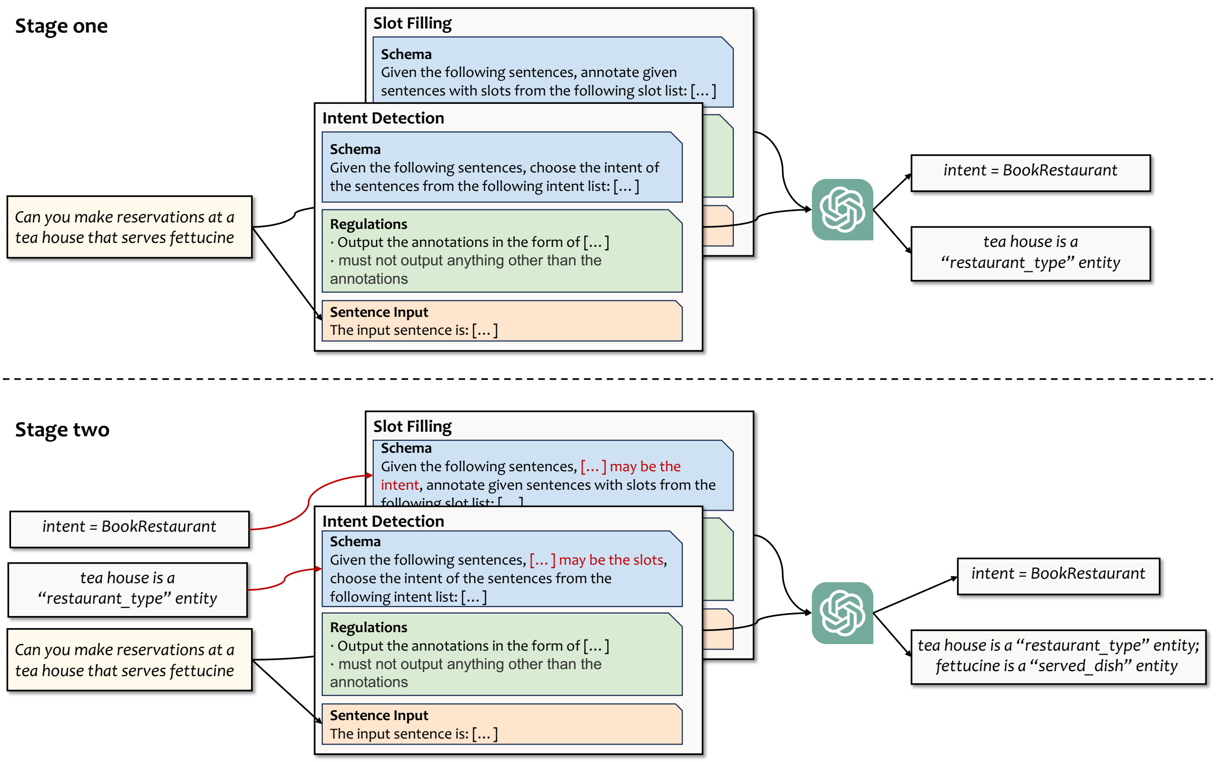

Figure 1: The overview for the proposed GPT-SLU framework. For illustration, we use the sample of SNIPS (Coucke et al., 2018) on two subtasks (intent section and slot filling).

SLURP (Bastianelli et al., 2020). Empirical results show that vanilla ChatGPT without using GPT-SLU achieves poor performance with original task instruction, while our two-stage framework based on ChatGPT achieves promising results.

2. Problem Definition

Given an utterance U, the task of SLU aims to output an intent label O^I and a slot label sequence $O^S = \{o_1^S, ..., o_n^S\}$, where n is the length of U.

3. GPT-SLU

We decompose the SLU task into two stages, each containing a single turn of QA, which refers to the dialogue with ChatGPT. The overview of the proposed GPT-SLU framework is shown in Figure 1, which we will describe in detail in the following.

3.1. Stage One

This stage generates initial intent and slots, which can be further decomposed into three components:

Schemas are designed to supply ChatGPT with crucial information to address SLU, guiding its generation process. They include intent constraints or slot constraints, which play a crucial role in accurately generating intents and slots. Specifically, the intent constraint is a comprehensive list of all possible intents available for ChatGPT, while the slot constraints offer examples of valid values and detailed descriptions associated with each slot.

Regulations are used to guide ChatGPT to generate reasonable responses. As shown in Figure 1, we require ChatGPT to first predict intent with template `"The intent is <intent>"`. Then, all extracted slot-value pairs are restricted in the form of `"<value> is an <slot> entity;..."`.

Input is the sample used for testing. Given the input in Figure 1 as an example, we ask ChatGPT to predict the corresponding intents and slots of sentence input *"Can you make reservations at a tea house that serves fettucine "*.

3.2. Stage Two

LLMs often suffer from the hallucination or overprediction issue (Ji et al., 2023; Wang et al., 2023). Moreover, the high correlations between the two tasks are not leveraged, which is a key aspect in previous supervised models. Therefore, we utilize intent and slots from stage one as cues to mutually guide each other, in a mutual verification manner.

Concretely, once the initial intent has been obtained, we incorporate this into the original statement and modify the schemas in stage one:

Model	SNIPS (Coucke et al., 2018)		ATIS (Hemphill et al., 1990)		SLURP (Bastianelli et al., 2020)	
	Intent (Acc)	Slot (F1)	Intent (Acc)	Slot (F1)	Intent (Acc)	Slot (F1)
Finetuned SOTA	99.12†	97.21†	98.54†	96.46†	85.26†	-
GPT-3.5 (*text-davinci-003*)	98.00*	68.90*	90.03*	55.72*	72.79	6.03
ChatGPT	97.71*	58.24*	75.22*	15.71*	-	-
GPT-SLU	98.50	75.65	88.90	67.04	80.21	18.75

Table 1: Results on three SLU benchmark datasets. "-" indicates the original paper does not report results. † denotes the results are obtained from corresponding papers Chen et al. (2022); Chang and Chen (2022). * denotes the results are cited from Pan et al. (2023).

[“intent/slots from stage one” may be the intent/slots]

for slot filling/intent detection task, respectively. In this manner, the prediction process of each subtask can be guided by the other, and the fruitful verification information from the other task also helps alleviate the issue of hallucinations.

4. Experiments

4.1. Datasets and Metrics

Dataset We use the test set of ATIS (Hemphill et al., 1990) and SNIPS (Coucke et al., 2018) to evaluate the zero-shot SLU performance. ATIS has 893 utterances for testing, while SNIPS has 700 ones for testing. To better fit the voice assistant application scenario, we also conduct experiments on SLURP (Bastianelli et al., 2020). SLURP is a large-scale dataset of commands to voice assistants with over 141k samples annotated with 60 different intents formulated as scenario-action pairs, as well as 56 types of entities or slots.

Metrics For metrics, we evaluate the performance of models on the widely-used SLU metrics (Goo et al., 2018), *i.e.*, accuracy (Acc) for intent detection and F1 score for slot filling.

4.2. Baselines

We compare our proposed GPT-SLU framework with the following baselines: (1) **GPT-3.5** (Brown et al., 2020; Ouyang et al., 2022) is a language model with 175B parameters that have been pre-trained on an extensive web corpus. In this paper, we use *text-davinci-003* version of GPT-3.5 from OpenAI API. (2) **ChatGPT** (Pan et al., 2023) is a ChatGPT-based method equipped with an in-context learning prompt template. (3) **state-of-the-art (SOTA) fine-tuned models** to provide a comparative analysis. Specifically, we choose the model proposed by Chen et al. (2022) on ATIS and SNIPS. On SLURP, we adopt the results used by Chang and Chen (2022).

4.3. Main Results

We report the main results in Table 1, from which we can draw the following conclusions: (1) While ChatGPT (Row ChatGPT, using vanilla single-stage prompt instead of GPT-SLU) performs poorly in solving SLU, our proposed two-stage framework based on ChatGPT (Row GPT-SLU) succeeds. GPT-SLU generally improves performance over three widely used SLU datasets significantly. (2) GPT-SLU surpasses GPT-3.5 on SNIPS and SLURP. We attribute it to the fact that the proposed multi-turn interactive prompts can better leverage ChatGPT's multi-turn ability to improve SLU performance. (3) The performance of ChatGPT on slot filling is significantly lower compared to intent detection. We intuitively suspect that this is due to the gap between the semantic labeling task and the text generation model leads to inferior performance when applying LLMs to resolve the slot filling task.

4.4. Model Analysis

Multi-stage mutual guided prompts can boost SLU To evaluate the effectiveness of the proposed two-stage framework, we employ a single-stage prompt to predict the SLU results. The results are shown in Table 2. We observe that the GPT-SLU surpasses the single-stage prompt across both metrics. We attribute it to the fact that multi-stage mutual guided prompts more effectively exploit the correlations between the two subtasks than the single-stage direct prompt, leveraging LLM's capabilities to harness the inter-task correlations.

Model	SLURP	
	Intent (Acc)	Slot (F1)
Two-stage mutual guided prompts	80.21	18.75
Single-stage prompt	75.59	13.35

Table 2: Results of prompt strategies on SLURP.

Utterance	is it going to be chillier at 10 pm in texas	**Utterance**	this textbook gets a two
Intent	GetWeather	**Intent**	The intent of the input sentence is RateBook
Slot	chillier: condition_temperature; at 10 pm: timeRange; texas: state	**Slot**	textbook: object_type; two: rating_value
	(a)		(b)

Figure 2: Two typical error cases of GPT-SLU.

Extra information can further facilitate SLU We evaluate the effectiveness of providing slot names only (Base), slot descriptions (w/ Des.), example (w/ Exp.), or a combination of the above following Pan et al. (2023). The results are presented in Table 3. We find that (1) both w/ Des. and W/ Exp. can provide extra information to boost performance; (2) the greater performance improvement of W/ Exp. compared to W/ Des. suggests that the model is better at learning the underlying mapping relationships through the provided samples; (3) providing both slot names and descriptions leads to the best performance of slot filling, indicating the importance of providing relevant information.

Model	SLURP	
	Intent (Acc)	Slot (F1)
Base	80.21	18.75
w/ Des.	80.85	19.54
w/ Exp.	81.03	19.68
w/ Des+Exp.	82.09	22.36

Table 3: Impact of Prompt Design on SLU Performance of GPT-SLU in stage one.

4.5. Error Analysis

Although our proposed GPT-SLU achieves promising results on three benchmark datasets, it still demonstrates some errors that may prevent the correct parsing of output. We summarize these errors into two main categories, which are shown in Figure 2: (1) **Format Violations**: Some outputs violate our format requirements. Take the prediction in Figure 2(a) as an example, GPT-SLU predicts `at 10 p.m.` as the value for slot `timeRange`, whereas the correct format for a time expression should not contain prepositions. (2) **Verbose Responses**: There are instances when GPT-SLU may generate natural language responses, even though we have implemented stringent constraints on the output. An example of a verbose output is illustrated in Figure 2(b). Therefore, it is necessary to perform post-processing on the output generated by GPT-SLU. An interesting direction is to explore integrating tools and plugins with GPT-SLU to enhance the standardization of SLU outputs.

5. Related Work

Spoken Language Understanding Spoken language understanding (SLU) is pivotal for accurately interpreting the user's intent through the construction of semantic frames (Qin et al., 2021). In general, SLU encompasses two subtasks: intent detection and slot filling. Due to the high correlations of the two subtasks, a bunch of models (Cheng et al., 2023c,d) have been proposed to tackle the two subtasks jointly. Due to the scarcity of data, a series of SLU models for more challenging scenarios have also been proposed, such as ASR-robust SLU (Cheng et al., 2023a), few-shot SLU (Hou et al., 2022), and zero-shot cross-lingual SLU (Zhu et al., 2023a) among others.

ChatGPT in NLP Application ChatGPT has gained widespread attention recently. Many fields received its impacts and evolving fast, such as Medicine (Jeblick et al., 2022) and Online Exam (Susnjak, 2022). In NLP, there are new investigations with ChatGPT in several tasks as well. For example, Zhang et al. (2022) use ChatGPT achieved state-of-the-art performance on Stance Detection, Guo et al. (2023) evaluated its helpfulness on question answering, Jiao et al. (2023a) state that it is a good translator for spoken language. Among them, Pan et al. (2023); He and Garner (2023); Li et al. (2023) first conducted a preliminary evaluation of ChatGPT for SLU tasks. We try to dig into its SLU ability, suggesting a two-stage mutual guided zero-shot SLU framework.

6. Conclusion

We presented GPT-SLU, a simple yet effective two-stage framework for zero-shot spoken language understanding (SLU) based on ChatGPT. Through the two-stage interactive mode, GPT-SLU facilitates mutual guidance and verification between the two subtasks, thereby mitigating errors and illusions to boost performance. We conducted experiments on three benchmark datasets to validate the effectiveness. Surprisingly, GPT-SLU achieves more impressive performance than its vanilla counterpart. We hope this work offers inspiration for zero-shot LLM-based spoken language understanding.

Limitations and Future Work There are several limitations in our GPT-SLU, which can be improved in future work: (1) The current multi-stage mutual guided prompt incurs a slightly higher cost. In future work, we will strive for single-step interaction to enable effective mutual guidance across multi-tasks. (2) It is also interesting to explore how LLMs can guide smaller supervised ones in SLU scenarios, which holds significant implications for practical voice assistant applications. (3) The evaluation benchmarks are limited, and the results may be sensitive to changes over time as versions of the LLMs are updated.

7. Bibliographical References

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. SLURP: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7252–7262. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ya-Hsin Chang and Yun-Nung Chen. 2022. Contrastive learning for improving ASR robustness in spoken language understanding. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 3458–3462. ISCA.

Dongsheng Chen, Zhiqi Huang, Xian Wu, Shen Ge, and Yuexian Zou. 2022. Towards joint intent detection and slot filling via higher-order attention. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4072–4078. ijcai.org.

Zhaorun Chen, Zhuokai Zhao, Zhihong Zhu, Ruiqi Zhang, Xiang Li, Bhiksha Raj, and Huaxiu Yao. 2024. Autoprm: Automating procedural supervision for multi-step reasoning via controllable question decomposition. *NAACL*.

Xuxin Cheng, Bowen Cao, Qichen Ye, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023a. ML-LMCL: Mutual learning and large-margin contrastive learning for improving ASR robustness in spoken language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6492–6505, Toronto, Canada. Association for Computational Linguistics.

Xuxin Cheng, Wanshi Xu, Ziyu Yao, Zhihong Zhu, Yaowei Li, Hongxiang Li, and Yuexian Zou. 2023b. FC-MTLF: A Fine- and Coarse-grained Multi-Task Learning Framework for Cross-Lingual Spoken Language Understanding. In *Proc. INTERSPEECH 2023*, pages 690–694.

Xuxin Cheng, Wanshi Xu, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023c. Towards spoken language understanding via multi-level multi-grained contrastive learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 326–336, New York, NY, USA. Association for Computing Machinery.

Xuxin Cheng, Zhihong Zhu, Bowen Cao, Qichen Ye, and Yuexian Zou. 2023d. MRRL: Modifying the reference via reinforcement learning for non-autoregressive joint multiple intent detection and slot filling. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10495–10505, Singapore. Association for Computational Linguistics.

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. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces.

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, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 753–757. Association for Computational Linguistics.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *CoRR*, abs/2301.07597.

Mutian He and Philip N. Garner. 2023. Can chatgpt detect intent? evaluating large language models for spoken language understanding. *CoRR*, abs/2305.13512.

Yutai Hou, Cheng Chen, Xianzhen Luo, Bohan Li, and Wanxiang Che. 2022. Inverse is better! fast and accurate prompt for few-shot slot tagging. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 637–647. Association for Computational Linguistics.

Yunpeng Huang, Yaonan Gu, Jingwei Xu, Zhihong Zhu, Zhaorun Chen, and Xiaoxing Ma. 2024. Securing reliability: A brief overview on enhancing in-context learning for foundation models. *arXiv preprint arXiv:2402.17671*.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian O. Sabel, Jens Ricke, and Michael Ingrisch. 2022. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *CoRR*, abs/2212.14882.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023a. Is chatgpt A good translator? A preliminary study. *CoRR*, abs/2301.08745.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. 2023b. Is chatgpt a good translator? yes with gpt-4 as the engine.

Guangpeng Li, Lu Chen, and Kai Yu. 2023. How ChatGPT is Robust for Spoken Language Understanding? In *Proc. INTERSPEECH 2023*, pages 2163–2167.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Wenbo Pan, Qiguang Chen, Xiao Xu, Wanxiang Che, and Libo Qin. 2023. A preliminary evaluation of chatgpt for zero-shot dialogue understanding. *CoRR*, abs/2304.04256.

Libo Qin, Qiguang Chen, Tianbao Xie, Qixin Li, Jian-Guang Lou, Wanxiang Che, and Min-Yen Kan. 2022. Gl-clef: A global-local contrastive learning framework for cross-lingual spoken language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2677–2686. Association for Computational Linguistics.

Libo Qin, Tianbao Xie, Wanxiang Che, and Ting Liu. 2021. A survey on spoken language understanding: Recent advances and new frontiers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4577–4584. ijcai.org.

Teo Susnjak. 2022. Chatgpt: The end of online exam integrity? *CoRR*, abs/2212.09292.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. GPT-NER: named entity recognition via large language models. *CoRR*, abs/2304.10428.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Ting-Wei Wu, Ruolin Su, and Biing-Hwang Juang. 2021. A label-aware BERT attention network for zero-shot multi-intent detection in spoken language understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 4884–4896. Association for Computational Linguistics.

Bowen Xing and Ivor W. Tsang. 2022. 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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 159–169. Association for Computational Linguistics.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081*.

Bowen Zhang, Daijun Ding, and Liwen Jing. 2022. How would stance detection techniques evolve after the launch of chatgpt? *CoRR*, abs/2212.14548.

Zhihong Zhu, Xuxin Cheng, Zhiqi Huang, Dongsheng Chen, and Yuexian Zou. 2023a. Enhancing code-switching for cross-lingual SLU: A unified view of semantic and grammatical coherence. In *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing, pages 7849–7856, Singapore. Association for Computational Linguistics.

Zhihong Zhu, Xuxin Cheng, Zhiqi Huang, Dongsheng Chen, and Yuexian Zou. 2023b. Towards unified spoken language understanding decoding via label-aware compact linguistics representations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12523–12531, Toronto, Canada. Association for Computational Linguistics.

Zhihong Zhu, Xuxin Cheng, Hongxiang Li, Yaowei Li, and Yuexian Zou. 2024. Dance with labels: Dual-heterogeneous label graph interaction for multi-intent spoken language understanding. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 1022–1031, New York, NY, USA. Association for Computing Machinery.

8. Language Resource References

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. SLURP: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7252–7262. Association for Computational Linguistics.

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. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.