A Survey on Spoken Language Understanding: Recent Advances and New Frontiers

Libo Qin, **Tianbao Xie**, **Wanxiang Che**[*], **Ting Liu**

Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
{lbqin, tianbaoxie, car, tliu}@ir.hit.edu.cn

Abstract

Spoken Language Understanding (SLU) aims to extract the semantics frame of user queries, which is a core component in a task-oriented dialog system. With the burst of deep neural networks and the evolution of pre-trained language models, the research of SLU has obtained significant breakthroughs. However, there remains a lack of a comprehensive survey summarizing existing approaches and recent trends, which motivated the work presented in this article. In this paper, we survey recent advances and new frontiers in SLU. Specifically, we give a thorough review of this research field, covering different aspects including (1) new taxonomy: we provide a new perspective for SLU filed, including *single model* vs. *joint model*, *implicit joint modeling* vs. *explicit joint modeling* in joint model, *non pre-trained paradigm* vs. *pre-trained paradigm*; (2) new frontiers: some emerging areas in complex SLU as well as the corresponding challenges; (3) abundant open-source resources: to help the community, we have collected, organized the related papers, baseline projects and leaderboard on a public website where SLU researchers could directly access to the recent progress. We hope that this survey can shed a light on future research in SLU field.

1 Introduction

Spoken Language Understanding (SLU) is a core component in task-oriented dialog system, which aims to capture the semantics of user queries. It typically consists of two tasks: intent detection and slot filling [Tur and De Mori, 2011]. Take the utterance *"I like to watch action movie"* in Figure 1 as an example, the outputs include an intent class label (i.e., `WatchMovie`) and a slot label sequence (i.e., `O, O, O, B-movie-type, I-movie-type, I-movie-type`).

Intent detection can be defined as a sentence classification problem. In recent years, many neural-network based classification methods such as convolutional neural network (CNN) [Xu and Sarikaya, 2013] and recurrent neural network(RNN) [Ravuri and Stolcke, 2015] have been investigated. Slot filling can be formulated as a sequence labeling task and popular sequence labeling methods such as conditional random field (CRF), RNN-based models [Xu and Sarikaya, 2013] and Long Short-Term Memory Network (LSTM) [Ravuri and Stolcke, 2015] have been explored.

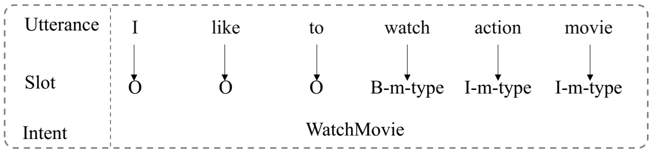
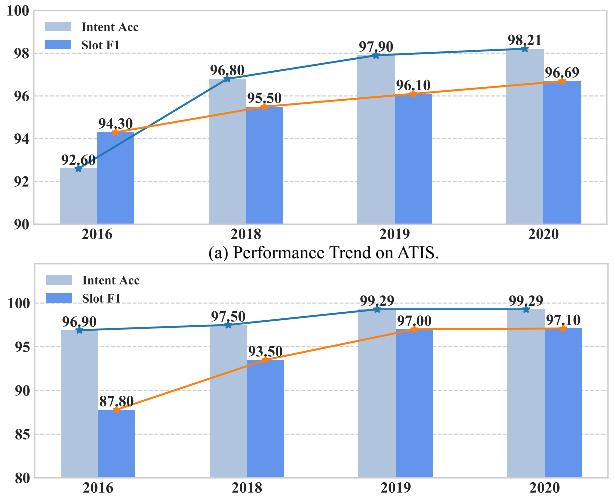


Figure 1: An example with intent and slot annotation (BIO format). m-type denotes movie-type.

Traditional approaches consider slot filling and intent detection as two separate tasks, which ignore the shared knowledge across the two tasks. Intuitively, intent detection and slot filling are not independent and highly tied. For example, if the intent of a user query is `WatchMovie`, it is more likely to contain the slot movie name rather than the slot music name. Thus, it's promising to consider the interaction between the two tasks. To this end, dominant models in the literature adopt joint models for leveraging shared knowledge across the two tasks, such as vanilla multi-task [Zhang and Wang, 2016], slot-gated [Goo *et al.*, 2018; Li *et al.*, 2018], stack-propagation [Qin *et al.*, 2019] and bi-directional interaction [E *et al.*, 2019; Qin *et al.*, 2021b]. With the popularity of deep learning and the emergence of pre-trained language models, SLU direction has made significant progress in recent years. As shown in Figure 2, in slot filling and intent detection tasks, we clearly observe that performance has even surpassed 97.0% and 98.0% on ATIS [Hemphill *et al.*, 1990] while 97% and 99% on SNIPS [Coucke *et al.*, 2018] that are the most wildly used datasets in SLU community. This leaves us with a question: *Have we achieved SLU tasks perfectly*?

In this paper, we introduce a survey to answer the above question including: 1) a comprehensive summary of recent progress in SLU field; 2) research challenges and frontiers for complex SLU tasks are concluded. Our survey observes that mainstream work remains the simple setting: single domain and single turn, which is still far from satisfying the requirements of some complex applications.

[*] Corresponding Author



(a) Performance Trend on ATIS.

(b) Performance Trend on SNIPS.

Figure 2: Recent Performance Trend.

In summary, the contributions of this survey can be concluded as follows:

- *New taxonomy*. We propose a taxonomy of SLU field, which categorizes existing approaches from three different perspectives: 1) *single model* vs. *joint model*; 2) *implicit joint model* vs. *explicit joint model* in joint model; 3) *non pre-trained models* vs. *pre-trained models*.
- *Abundant resources*. We collect abundant resources on SLU including open-source implementations, corpora, and paper lists[1]. To our knowledge, this is the first effort to collect open-source resources for SLU community.
- *New Frontiers*. We discuss and analyze the limitations of existing SLU. Also, we suggest some new frontiers and discuss the challenges.

We hope this survey will help researchers to understand the latest progress, challenges and frontiers in SLU field.

The rest of the survey is organized as follows. Section 2 outlines the background of SLU. Section 3 gives a brief overview proposed taxonomy of SLU. Section 4 discusses the new frontiers and their challenges. Section 5 gives the related survey on SLU. Section 6 summarizes the paper.

2 Background

In this section, we describe the definition for slot filling, intent detection and joint model, and then we give a brief description of the wildly used datasets and evaluation metrics.

2.1 Definition

Intent Detection. Given input utterance $X = (x_1, \ldots, x_n)$ (n denotes the length of X), intent detection (ID) can be considered as a sentence classification task to decide the intent label o^I, which is formulated as:

$$o^I = \texttt{Intent-Detection}(X). \tag{1}$$

Model	Intent Acc	Slot F1
Bi-Jordan RNN [Mesnil *et al.*, 2013]	-	93.98
RNN [Yao *et al.*, 2013]	-	94.11
Hybrid RNN [Mesnil *et al.*, 2014]	-	95.06
LSTM [Yao *et al.*, 2014a]	-	95.08
R-CRF [Yao *et al.*, 2014b]	-	96.65
RNN [Ravuri and Stolcke, 2015]	97.55	-
LSTM [Ravuri and Stolcke, 2015]	98.06	-
RNN SOP [Liu and Lane, 2015]	-	94.89
5xR-biRNN [Vu *et al.*, 2016]	-	95.56
Encoder-labeler [Kurata *et al.*, 2016]	-	95.66

Table 1: Single model performance on intent detection and slot filling on ATIS. Acc denotes the accuracy metric.

Slot Filling. Slot filling (SF) can be seen as a sequence labeling task to produce a sequence slots $o^S = (o^S_1, \ldots, o^S_n)$, which can be written as:

$$o^S = \texttt{Slot-Filling}(X). \tag{2}$$

Joint Model. Joint model denotes that a joint model predicts the slots sequence and intent simultaneously, which has the advantage of capturing shared knowledge across related tasks, using:

$$(o^I, o^S) = \texttt{Joint-Model}(X). \tag{3}$$

2.2 Dataset

The most wildly used datasets are ATIS [Hemphill *et al.*, 1990] and SNIPS [Coucke *et al.*, 2018]. In the following, we will give a detailed description.

ATIS. ATIS dataset contains audio recordings of flight, reservations. There are 4,478 utterances for training, 500 utterances for validation and 893 utterances for testing. 120 slot labels and 21 intent types are included in ATIS training data.

SNIPS. SNIPS is the custom-intent-engines collected by Snips [Coucke *et al.*, 2018]. There are 13,084 utterances for training, 700 utterances for validation and 700 utterances for testing. There are 72 slot labels and 7 intent types.

2.3 Evaluation Metrics

The most wildly used evaluation metrics for SLU are F1 scores, intent accuracy and overall accuracy.

- **F1 scores:** F1 scores are adopted to evaluate the performance of slot filling, which is the harmonic mean score between precision and recall. A slot prediction is considered correct when an exact match is found [Tjong Kim Sang and De Meulder, 2003].
- **Intent Accuracy:** Intent Accuracy is used for evaluating the performance of intent detection, calculating the ratio of sentences for which intent is predicted correctly.
- **Overall Accuracy:** Overall accuracy is adopted for calculating the ratio of sentences for which both intent and slot are predicted correctly in a sentence [Goo *et al.*, 2018]. This metric considers intent detection and slot filling simultaneously.

[1] https://github.com/yizhen20133868/Awesome-SLU-Survey

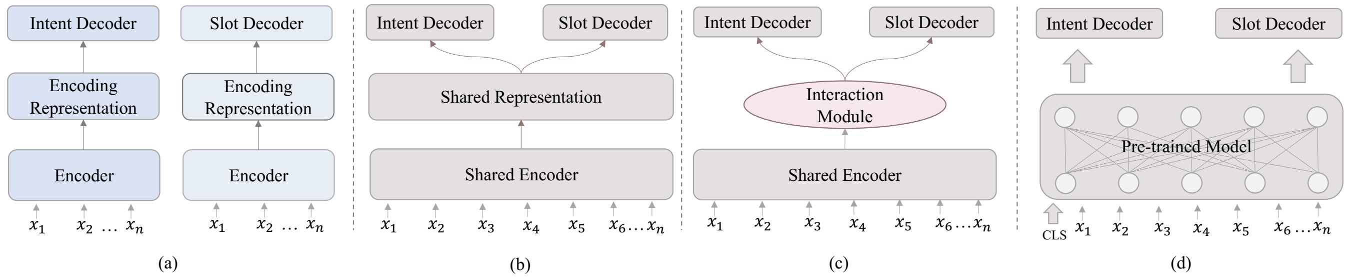


Figure 3: (a) Single models. (b) Implicit Joint Modeling. (c) Explicit Joint Modeling. (d) Pre-trained Model Paradigm.

3 Taxonomy

In this section, we describe the taxonomy in SLU, including *single models* (§3.1), *joint models* (§3.2) and *pre-trained paradigm* (§3.3), which is shown in Figure 3.

3.1 Single Model

Single models train each task *separately* for intent detection and slot filling, which is shown in Figure 3(a).

Intent Detection. Many sentence classification methods have been investigated for intent detection. Xu and Sarikaya [2013] utilized Convolutional Neural Network (CNN) to extract 5-gram features and apply max-pooling to obtain word representations. Ravuri and Stolcke [2015] successfully applied Long Short-Term Memory Network(LSTM) to the ID task, which indicates the sequential features are beneficial to intent detection task.

Slot Filling. Popular neural approaches for slot filling including CRF, Recurrent neural network (RNN) and RNN-based models. Yao *et al.* [2013] adopted RNN Language Models (RNN-LMs) to predict slot labels rather than words. In addition, RNN-LMs explored future words, named entities, syntactic features and word-class information. Mesnil *et al.* [2013] investigated several RNN architectures including Elman RNN, Jordan RNN and its bi-directional version for SLU. Yao *et al.* [2014a] proposed an LSTM framework for the slot filling task. Mesnil *et al.* [2014] applied Viterbi encodings and recurrent CRFs to eliminate the label bias problem. Yao *et al.* [2014b] proposed R-CRF to tackle label bias. Liu and Lane [2015] proposed to model slot label dependencies using a sampling approach, by feeding sampled output labels (true or predicted) back to the sequence state. Vu *et al.* [2016] utilized ranking loss function on bi-RNN model in SF, further enhancing performance in ATIS dataset. Kurata *et al.* [2016] leveraged sentence-level information from encoder to improve performance for SF task.

The wildly used dataset for evaluating single models is ATIS. Table 1 summarizes the performance of single model on intent detection and slot filling.

Highlight. There is no interaction between intent detection and slot filling in single models due to the separate training, leading to shared knowledge leakage across two tasks.

3.2 Joint Model

Considering the close correlation between intent detection and slot filling, dominant work in the literature adopts joint model to leverage the shared knowledge across tasks. Existing joint work can be classified into two main categories: *implicit joint modeling* and *explicit joint modeling*.

Implicit Joint Modeling. Implicit joint modeling denotes that model only adopts a shared encoder to capture shared features, without any explicit interaction, which is illustrated in Figure 3(b). Zhang and Wang [2016] introduced a shared RNNs (Joint ID and SF) to learn the correlation between intent and slots. Liu and Lane [2016a] introduced a shared encoder decoder framework with attention-mechanism (Attention BiRNN) for intent detection and slot filling. Liu and Lane [2016b] used a shared RNN to jointly perform SF, ID and language modeling (Joint SLU-LM), aiming to improve the ability of online prediction. Hakkani-Tür *et al.* [2016] proposed a shared RNN-LSTM architecture (Joint Seq) for joint modeling.

Highlight. Though *implicit joint modeling* is a direct method to incorporate the shared knowledge, it does not model the interaction explicitly, resulting in low interpretability and low performance.

Explicit Joint modeling. In recent years, more and more work has been proposed to explicitly model the interaction between intent detection and slot filling with an explicit interaction module, which is shown in Figure 3(c). This explicit modeling mode has the advantages of explicitly controlling process of interaction. The existing *explicit joint modeling* methods can be categorized into two types: *single flow interaction* and *bidirectional flow interaction*.

- ***Single Flow Interaction:*** Recent work on *single flow interaction* mainly considered the single information flow from intent to slot. Goo *et al.* [2018] proposed a slot-gated joint model (Slot-Gated), which allows the slot filling be can be conditioned on the learned intent. Li *et al.* [2018] proposed a novel self-attentive model (Self-Atten. Model) with the intent augmented gate mechanism to guide the slot filling. Qin *et al.* [2019] proposed a stack-propagation model to directly use intent detection results to guide slot filling and used the token-level intent detection to alleviate the error propagation, further enhancing the performance.
- ***Bidirectional Flow Interaction:*** Bidirectional flow interaction work means that model considered the cross-impact between intent detection and slot filling. As exploration, Wang *et al.* [2018] proposed a Bi-Model architecture to consider the cross-impact across SF and

	ATIS			SNIPS		
Model	Intent Acc	Slot F1	Overall Acc	Intent Acc	Slot F1	Overall Acc
Implicit Joint Modeling						
Joint ID and SF [Zhang and Wang, 2016]	98.32	96.89	-	-	-	-
Attention BiRNN [Liu and Lane, 2016a]	91.10	94.20	78.90	96.70	87.80	74.10
Joint SLU-LM [Liu and Lane, 2016b]	98.43	94.47	-	-	-	-
Joint Seq. [Hakkani-Tür *et al.*, 2016]	92.60	94.30	80.70	96.90	87.30	73.20
Explicit Joint Modeling						
Slot-Gated [Goo *et al.*, 2018]	93.60	94.80	82.20	97.00	88.80	75.50
Self-Atten. Model [Li *et al.*, 2018]	96.80	95.10	82.20	97.50	90.00	81.00
Bi-model [Wang *et al.*, 2018]	96.40	95.50	85.70	97.20	93.50	83.80
SF-ID Network [E *et al.*, 2019]	97.09	95.80	86.90	97.29	92.23	80.43
Capsule-NLU [Zhang *et al.*, 2019]	95.00	95.20	83.40	97.30	91.80	80.90
CM-Net [Liu *et al.*, 2019]	96.10	95.60	85.30	98.00	93.40	84.10
Stack-Propgation [Qin *et al.*, 2019]	96.90	95.90	86.50	98.00	94.20	86.90
Graph LSTM [Zhang *et al.*, 2020b]	97.20	95.91	87.57	98.29	95.30	89.71
Co-Interactive transformer [Qin *et al.*, 2021b]	97.70	95.90	87.40	98.80	95.90	90.30
Pre-trained Models						
BERT-Joint [Castellucci *et al.*, 2019]	97.80	95.70	88.20	99.00	96.20	91.60
Joint BERT +CRF [Chen *et al.*, 2019]	97.90	96.00	88.60	98.40	96.70	92.60
Stack-Propgation +BERT [Qin *et al.*, 2019]	97.50	96.10	88.60	99.00	97.00	92.90
Co-Interactive transformer +BERT [Qin *et al.*, 2021b]	98.00	96.10	88.80	98.80	97.10	93.10

Table 2: Joint model performance on intent detection and slot filling. Acc denotes the accuracy metric. We adopted reported results from published literature [Goo *et al.*, 2018] and [Qin *et al.*, 2021b].

ID by using two correlated bidirectional LSTMs. E *et al.* [2019] proposed a novel SF-ID network that provides a bi-directional interrelated mechanism for SF and ID tasks, considering the influence of SF-to-ID and ID-to-SF. Zhang *et al.* [2019] introduced a dynamic routing capsule network (Capsule-NLU) to incorporate hierarchical and interrelated relationships among two tasks. Liu *et al.* [2019] proposed a novel collaborative memory network (CM-Net) for jointly modeling SF and ID. Zhang *et al.* [2020b] do exploration in introducing graph LSTM to SLU, achieving the promising performance. Qin *et al.* [2021b] proposed a co-interactive transformer to consider the cross-impact by building a bidirectional connection between the two related tasks.

The wildly used datasets for evaluating joint models are ATIS and SNIPS. Table 2 concludes the performance.

Highlight. Compared with *implicit joint modeling* method, *explicit joint modeling* has the following advantages. First, a simple multi-task framework just implicitly considers mutual connection between two tasks by sharing latent representations, which cannot achieve desirable results. In contrast, *explicit joint modeling* can enable model to fully capture the shared knowledge across tasks, which promotes the performance on two tasks. Second, explicitly controlling knowledge transfer for two tasks can help to improve interpretability where impact between SF and ID can be analyzed easily.

3.3 Pre-trained Paradigm

Recently, Pre-trained Language Models (PLMs) achieve surprising results across various NLP tasks [Wang *et al.*, 2020]. Some BERT-based [Devlin *et al.*, 2019] pre-trained work has been explored in SLU direction where a shared BERT is considered as the encoder to extract contextual representations. In BERT-based models, each utterance starts with `[CLS]` and ends with `[SEP]`, where `[CLS]` is the special symbol for representing the whole sequence, and `[SEP]` is the special symbol to separate non-consecutive token sequences. Further, the representation of the special token `[CLS]` is used for intent detection while other token representations are adopted for slot filling, which is shown in Figure 3(d).

More specifically, Chen *et al.* [2019] explored BERT for SLU where BERT is used to extract shared contextual embedding for intent detection and slot filling, which obtains a significant improvement compared with other non pre-trained models. Castellucci *et al.* [2019] used the simlilar architecture (BERT-Joint) for jointly modeling intent detection and slot filing. Qin *et al.* [2019] used pre-trained embedding encoder to replace its attention encoder (Stack-Propgation +BERT), further boosting model's performance. Qin *et al.* [2021b] also explored BERT for SLU (Co-Interactive transformer +BERT), which achieves state-of-the-art performance. Table 2 shows the results of pre-trained models.

Highlight. Pre-trained models can provide rich semantic features, which can help to improve the performance on SLU tasks. This observation is consistent with pre-trained models for other NLP applications.

4 New Frontiers and Challenges

Section 3 has discussed the traditional SLU setting that mainly focuses on the single-domain or single-turn setting, which limits its application and may be not enough for complex needs in the real-world scenario. In the following, we will discuss new frontiers in a complex setting and their challenges, including *contextual SLU* (§4.1), *multi-intent SLU* (§4.2), *Chinese SLU* (§4.3), *cross-domain SLU* (§4.4), *cross-*

lingual SLU (§4.5) and *low-resource SLU* (§4.6).

4.1 Contextual SLU

Naturally, completing a task usually necessitates multiple turns of back-and-forth conversations between the user and the system, which requires model to consider the contextual SLU. Unlike the single turn SLU, contextual SLU faces unique ambiguity challenge, since the user and the system may refer to entities introduced in prior dialogue turns, introducing ambiguity, which requires model to incorporate the contextual information for alleviating ambiguity.

To this end, Chen *et al.* [2016] proposed a memory network to incorporate dialogue history information, showing that their model outperforms models without context. Bapna *et al.* [2017a] proposed a sequential dialogue encoder network, which allows encoding context from the dialogue history in chronological order. Su *et al.* [2018] designed and investigated various time-decay attention functions based on an end-to-end contextual language understanding model. Qin *et al.* [2021a] proposed an adaptive fusion layer to dynamically consider the different and relevant contextual information for guiding the slot filling, achieving a fine-grained contextual information transfer.

The main challenges for contextual SLU are as follows: (1) **Contextual Information Integration:** Correctly differentiating relevance between different dialog histories with the current utterance and effectively incorporating contextual information into contextual SLU is a core challenge. (2) **Long Distance Obstacles:** Since some dialogues have extreme long histories, how to effectively model the long-distance dialog history and filter irrelevant noise is an interesting research topic.

4.2 Multi-Intent SLU

Multi-intent SLU means that the system can handle an utterance containing multiple intents and its corresponding slots. Gangadharaiah and Narayanaswamy [2019] show that 52% of examples are multi-intent in the amazon internal dataset, which indicates that multi-intent setting is more practical in the real-world scenario.

To this end, Gangadharaiah and Narayanaswamy [2019] explored a multi-task framework to jointly perform multi-intent classification and slot filling. Qin *et al.* [2020c] proposed an adaptive graph-interactive framework to model the interaction between multiple intents and slot at each token.

There are several possible reasons for the slow progress of multi-intent SLU: (1) **Interaction between Multiple Intents and Slots:** Unlike the single intent SLU, how to effectively incorporate multiple intents information to lead the slot prediction is a unique challenge in multi-intent SLU. (2) **Lack of Data:** There is no human-annotated data for multi-intent SLU yet, which is another possible reason for the slow progress.

4.3 Chinese SLU

Chinese SLU means that an SLU model trained on Chinese data and directly be applied to Chinese community. Compared with SLU in English, Chinese SLU faces a unique challenge since it usually needs word segmentation.

Liu *et al.* [2019] contributed a new corpus (CAIS) to the research community. In addition, they proposed a character-based joint model to perform Chinese SLU. Though achieving promising performance, one drawback of the character-based SLU model is that explicit word sequence information is not fully exploited, which can be potentially useful. Teng *et al.* [2021] proposed a multi-level word adapter to effectively incorporate word information for both *sentence-level* intent detection and *token-level* slot filling.

The main challenges for Chinese SLU are as follows: (1) **Word Information Integration:** How to effectively incorporate word information to guide Chinese SLU is a unique challenge; (2) **Multiple Word Segmentation Criteria:** Since there are multiple word segmentation criteria, how to effectively combine multiple word segmentation information for Chinese SLU is non-trivial.

4.4 Cross Domain SLU

Though achieving promising performance in single domain setting, the existing SLU models rely on a considerable amount of annotated data, which limits their usefulness for new and extended domains. In practice, it's infeasible to collect rich labeled datasets for each new domain. Hence, it's promising to consider the cross-domain setting.

Methods in this area can be concluded into two categories: *Implicit domain knowledge transfer* with parameters sharing and *explicit domain knowledge transfer*. *Implicit domain knowledge transfer* means that model simply combines multi-domain datasets for training to capture domain features. Hakkani-Tür *et al.* [2016] proposed a single LSTM model over mixed multi-domain dataset, which can implicitly learn the domain-shared features. Kim *et al.* [2017a] adopted one network to jointly modeling slot filling, intent detection and domain classification, implicitly learning the domain-shared and task-shared information. Such methods can implicitly extract the shared features but fail to effectively capture domain-specific features.

Explicit domain knowledge transfer approaches denote that models used shared-private framework including a shared module to capture domain-shared feature and a private module for each domain, which has the advantage of explicitly differentiating shared and private knowledge. Kim *et al.* [2017b] used attention mechanism to learn a weighted combinations from the feedback of the expert models on different domains. Liu and Lane [2017] also used a shared LSTM to capture domain-shared knowledge and private LSTM to extract domain-specific feature, combining them for multi-domain slot filling. Qin *et al.* [2020b] proposed a model with separate domain- and task-specific parameters, which enables model to capture the task-aware and domain-aware features for multi-domain SLU.

Although there are many methods of cross-domain SLU, main challenges are as follows: (1) **Domain Knowledge Transfer:** Transferring knowledge from source domain to target domain is non-trivial. (2) **Zero-shot Setting:** When the target domain has no training data, how to transfer knowledge from source domain data to target domain is a challenge.

4.5 Cross-Lingual SLU

Cross-lingual SLU means that an SLU system trained on English can be directly applied to other low-resource languages, which has attracted more and more attention.

In recent years, Schuster *et al.* [2019] released a multilingual SLU dataset[2] contains English, Spanish and Thai to facilitate the cross-lingual SLU direction. Liu *et al.* [2020a] proposed an attention-Informed mixed-language training to align representation between source language and target language. Xu *et al.* [2020] proposed a novel end-to-end model that learns to align and predict target slot labels jointly for cross-lingual transfer. Besides, they introduced MultiATIS++, a new multilingual NLU corpus to the community. Qin *et al.* [2020a] proposed a multilingual code-switching augmentation framework to fine-tune mBERT for aligning representations from source and multiple target languages, which achieved promising performance.

Cross-lingual SLU has many interesting problems to focus on: (1) **Alignment between Source and Target Language:** Aligning intent and slot representations from source and target languages is a core challenge for cross-lingual SLU. (2) **Generalizability:** Since new languages emerge frequently, training a generalized SLU model that can applied to all languages deserves to be explored.

4.6 Low-resource SLU

The remarkable progress on SLU heavily relies on a considerable amount of labeled training data, which fails to work in low-resources setting where few or zero shot data can be accessed to. In this section, we will discuss the trends and progress on low-resource SLU, including *Few-shot SLU*, *Zero-shot SLU*, *Unsupervised SLU* and *Open-set SLU*.

Few-shot SLU. In some cases, a slot or intent only has fewer instances, which makes the traditional supervised SLU models powerless. To alleviate this issue, *few-shot SLU* is appealing in this scenario since it is able to adapt to new application quickly with only very few examples.

Recently, some work has been proposed for investigating this direction. Hou *et al.* [2020] proposed a few-shot CRF model with collapsed dependency transfer mechanism for few-shot slot tagging. Besides, Hou *et al.* [2021] have began to explore the few-shot multi-intent detection.

Zero-shot SLU. Face the rapid changing of applications, situations like no target training data may happen in a brand new application. Many zero shot methods offer a way of solving it by discovering commonalities between slots. Bapna *et al.* [2017b] proposed a method utilized slot description to obtain and transfer concept through different applications, empowering the models' zero-shot ability. Liu *et al.* [2020b] followed similar architecture, which trains the model on awareness of slot descriptions. Shah *et al.* [2019] countered the problems of misaligned overlapping schemas by adding slot examples values along with descriptions during training.

Unsupervised SLU. In recent years, unsupervised method has been proposed to automatically extract slot-value pairs, which is a promising direction to free model from heavy human annotation. To this end, Min *et al.* [2020] proposed a new task named *dialogue state induction*, which is to automatically identify dialogue state slot-value pairs.

Open-set SLU. Researchers also extended this problem to a more challenging open setting in recent years. Zhang *et al.* [2021] proposed to use known intents as prior knowledge to detect the one-class open intent during testing. Furthermore, they used clustering technologies to discover fine-grained new intent classes [Zhang *et al.*, 2020a; Lin *et al.*, 2020].

Low-resource SLU has attracted more and more attention and the main challenges are as follows: (1) **Interaction on low-resource setting:** How to make full use of connection between intents and slots in the low-resource setting is still under-explored. (2) **Lack of Benchmark:** There is lacking of public benchmarks in low-resource setting, which may impede the progress.

5 Related Work

Tur and De Mori [2011] gave a summary of the SLU field at the advent of the neural era (2011). Chen *et al.* [2017] provided a survey on task-oriented dialog systems, covering a small overall of state-of-the-art SLU models. Louvan and Magnini [2020] provided a good survey of intent detection and slot filling methods up to 2019.

Compared with their work, the main differences are summarized as follows: (1) In this survey, we introduce a new taxonomy of the technical architecture of SLU and we conduct a comprehensive review of the origin and the development of SLU; (2) We discuss and analyze the limitation of existing SLU and shed light on some new trends and discuss new frontiers in this research field. (3) We provide an open-sourced website for SLU researchers, hoping to facilitate the SLU field. We hope that this survey can shed a light on future research in SLU community.

6 Conclusion

This article presented a comprehensive survey on the progress of spoken language understanding. Based on a thorough analysis of recent work, we presented a new taxonomy of SLU from different modeling perspectives. In addition, considering the limitations of recent SLU system, we shed light on some new trends and discuss new frontiers in this research field. Finally, we made the first attempt to provide an open-sourced website including SLU datasets, papers, baseline projects and leaderboards for SLU researchers, hoping to facilitate the SLU community.

Acknowledgements

Thanks to Yudi Zhang and Lehan Wang for their helpful discussion and contribution. This work was supported by the National Key R&D Program of China via grant 2020AAA0106501 and the National Natural Science Foundation of China (NSFC) via grant 61976072 and 61772153. This work was supported by the Zhejiang Lab's International Talent Fund for Young Professionals.

[2] https://fb.me/multilingual_task_oriented_data

References

[Bapna *et al.*, 2017a] Ankur Bapna, Gokhan Tür, Dilek Hakkani-Tür, and Larry Heck. Sequential dialogue context modeling for spoken language understanding. In *Proc. of SIGdial*, 2017.

[Bapna *et al.*, 2017b] Ankur Bapna, Gokhan Tür, Dilek Hakkani-Tür, and Larry Heck. Towards zero-shot frame semantic parsing for domain scaling. In *Interspeech*, 2017.

[Castellucci *et al.*, 2019] Giuseppe Castellucci, Valentina Bellomaria, Andrea Favalli, and Raniero Romagnoli. Multi-lingual intent detection and slot filling in a joint bert-based model. *arXiv preprint arXiv:1907.02884*, 2019.

[Chen *et al.*, 2016] Yun-Nung Vivian Chen, Dilek Hakkani-Tür, Gokhan Tur, Jianfeng Gao, and Li Deng. End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding. In *Interspeech*, 2016.

[Chen *et al.*, 2017] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGkdd Explorations*, 2017.

[Chen *et al.*, 2019] Qian Chen, Zhu Zhuo, and Wen Wang. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*, 2019.

[Coucke *et al.*, 2018] Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*, 2018.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, 2019.

[E *et al.*, 2019] Haihong E, Peiqing Niu, Zhongfu Chen, and Meina Song. A novel bi-directional interrelated model for joint intent detection and slot filling. In *Proc. of ACL*, 2019.

[Gangadharaiah and Narayanaswamy, 2019] Rashmi Gangadharaiah and Balakrishnan Narayanaswamy. Joint multiple intent detection and slot labeling for goal-oriented dialog. In *Proc. of NAACL*, 2019.

[Goo *et al.*, 2018] Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. Slot-gated modeling for joint slot filling and intent prediction. In *Proc. of NAACL*, 2018.

[Hakkani-Tür *et al.*, 2016] Dilek Hakkani-Tür, Gökhan Tür, Asli Celikyilmaz, Yun-Nung Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. In *Interspeech*, 2016.

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

[Hou *et al.*, 2020] Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. In *Proc. of ACL*, 2020.

[Hou *et al.*, 2021] Yutai Hou, Yongkui Lai, Yushan Wu, Wanxiang Che, and Ting Liu. Few-shot learning for multi-label intent detection. In *Proc. of AAAI*, 2021.

[Kim *et al.*, 2017a] Young-Bum Kim, Sungjin Lee, and Karl Stratos. Onenet: Joint domain, intent, slot prediction for spoken language understanding. In *ASRU*, 2017.

[Kim *et al.*, 2017b] Young-Bum Kim, Karl Stratos, and Dongchan Kim. Domain attention with an ensemble of experts. In *Proc. of ACL*, 2017.

[Kurata *et al.*, 2016] Gakuto Kurata, Bing Xiang, Bowen Zhou, and Mo Yu. Leveraging sentence-level information with encoder LSTM for semantic slot filling. In *Proc. of EMNLP*, 2016.

[Li *et al.*, 2018] Changliang Li, Liang Li, and Ji Qi. A self-attentive model with gate mechanism for spoken language understanding. In *Proc. of EMNLP*, 2018.

[Lin *et al.*, 2020] Ting-En Lin, Hua Xu, and Hanlei Zhang. Discovering new intents via constrained deep adaptive clustering with cluster refinement. In *Proc of AAAI*, 2020.

[Liu and Lane, 2015] Bing Liu and Ian Lane. Recurrent neural network structured output prediction for spoken language understanding. In *Proc. of NIPS*, 2015.

[Liu and Lane, 2016a] Bing Liu and Ian Lane. Attention-based recurrent neural network models for joint intent detection and slot filling. In *Interspeech*, 2016.

[Liu and Lane, 2016b] Bing Liu and Ian Lane. Joint online spoken language understanding and language modeling with recurrent neural networks. In *Proc. of SIGdial*, 2016.

[Liu and Lane, 2017] Bing Liu and Ian Lane. Multi-domain adversarial learning for slot filling in spoken language understanding. In *NIPS Workshop*, 2017.

[Liu *et al.*, 2019] Yijin Liu, Fandong Meng, Jinchao Zhang, Jie Zhou, Yufeng Chen, and Jinan Xu. CM-net: A novel collaborative memory network for spoken language understanding. In *Proc. of EMNLP-IJCNLP*, 2019.

[Liu *et al.*, 2020a] Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proc. of AAAI*, 2020.

[Liu *et al.*, 2020b] Zihan Liu, Genta Indra Winata, Peng Xu, and Pascale Fung. Coach: A coarse-to-fine approach for cross-domain slot filling. In *Proc. of ACL*, 2020.

[Louvan and Magnini, 2020] Samuel Louvan and Bernardo Magnini. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. In *Proc. of COLING*, 2020.

[Mesnil *et al.*, 2013] Grégoire Mesnil, Xiaodong He, Li Deng, and Yoshua Bengio. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *Interspeech*, 2013.

[Mesnil *et al.*, 2014] Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, et al. Using recurrent neural networks for slot filling in spoken language understanding. *TASLP*, 2014.

[Min *et al.*, 2020] Qingkai Min, Libo Qin, Zhiyang Teng, Xiao Liu, and Yue Zhang. Dialogue state induction using neural latent variable models. In *Proc. of IJCAI*, 2020.

[Qin *et al.*, 2019] 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 *Proc. of EMNLP-IJCNLP*, 2019.

[Qin *et al.*, 2020a] Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. In *Proc. of IJCAI*, 2020.

[Qin *et al.*, 2020b] Libo Qin, Minheng Ni, Yue Zhang, Wanxiang Che, Yangming Li, and Ting Liu. Multi-domain spoken language understanding using domain-and task-aware parameterization. *arXiv preprint arXiv:2004.14871*, 2020.

[Qin *et al.*, 2020c] Libo Qin, Xiao Xu, Wanxiang Che, and Ting Liu. AGIF: An adaptive graph-interactive framework for joint multiple intent detection and slot filling. In *EMNLP Findings*, 2020.

[Qin *et al.*, 2021a] L. Qin, W. Che, M. Ni, Y. Li, and T. Liu. Knowing where to leverage: Context-aware graph convolution network with an adaptive fusion layer for contextual spoken language understanding. *TASLP*, 2021.

[Qin *et al.*, 2021b] 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.

[Ravuri and Stolcke, 2015] Suman Ravuri and Andreas Stolcke. Recurrent neural network and lstm models for lexical utterance classification. In *Interspeech*, 2015.

[Schuster *et al.*, 2019] Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proc. of NAACL*, 2019.

[Shah *et al.*, 2019] Darsh Shah, Raghav Gupta, Amir Fayazi, and Dilek Hakkani-Tur. Robust zero-shot cross-domain slot filling with example values. In *Proc. of ACL*, 2019.

[Su *et al.*, 2018] Shang-Yu Su, Pei-Chieh Yuan, and Yun-Nung Chen. How time matters: Learning time-decay attention for contextual spoken language understanding in dialogues. In *Proc. of NAACL*, 2018.

[Teng *et al.*, 2021] Dechuang Teng, Libo Qin, Wanxiang Che, Sendong Zhao, and Ting Liu. Injecting word information with multi-level word adapter for chinese spoken language understanding. In *ICASSP*, 2021.

[Tjong Kim Sang and De Meulder, 2003] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proc. of HLT-NAACL*, 2003.

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

[Vu *et al.*, 2016] N. T. Vu, P. Gupta, H. Adel, and H. Schütze. Bi-directional recurrent neural network with ranking loss for spoken language understanding. In *ICASSP*, 2016.

[Wang *et al.*, 2018] Yu Wang, Yilin Shen, and Hongxia Jin. A bi-model based RNN semantic frame parsing model for intent detection and slot filling. In *Proc. of NAACL*, 2018.

[Wang *et al.*, 2020] Yuxuan Wang, Yutai Hou, Wanxiang Che, and Ting Liu. From static to dynamic word representations: a survey. *IJMLC*, 2020.

[Xu and Sarikaya, 2013] Puyang Xu and Ruhi Sarikaya. Convolutional neural network based triangular crf for joint intent detection and slot filling. In *ASRU*, 2013.

[Xu *et al.*, 2020] Weijia Xu, Batool Haider, and Saab Mansour. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proc. of EMNLP*, 2020.

[Yao *et al.*, 2013] Kaisheng Yao, Geoffrey Zweig, Mei-Yuh Hwang, Yangyang Shi, and Dong Yu. Recurrent neural networks for language understanding. In *Interspeech*, 2013.

[Yao *et al.*, 2014a] Kaisheng Yao, Baolin Peng, Yu Zhang, Dong Yu, Geoffrey Zweig, and Yangyang Shi. Spoken language understanding using long short-term memory neural networks. In *SLT*, 2014.

[Yao *et al.*, 2014b] Kaisheng Yao, Baolin Peng, Geoffrey Zweig, Dong Yu, Xiaolong Li, and Feng Gao. Recurrent conditional random field for language understanding. In *ICASSP*, 2014.

[Zhang and Wang, 2016] Xiaodong Zhang and Houfeng Wang. A joint model of intent determination and slot filling for spoken language understanding. In *Proc. of IJCAI*, 2016.

[Zhang *et al.*, 2019] Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, and Philip Yu. Joint slot filling and intent detection via capsule neural networks. In *Proc. of ACL*, 2019.

[Zhang *et al.*, 2020a] Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lv. Discovering new intents with deep aligned clustering. *arXiv preprint arXiv:2012.08987*, 2020.

[Zhang *et al.*, 2020b] Linhao Zhang, Dehong Ma, Xiaodong Zhang, Xiaohui Yan, and Houfeng Wang. Graph lstm with context-gated mechanism for spoken language understanding. In *Proc. of AAAI*, 2020.

[Zhang *et al.*, 2021] Hanlei Zhang, Hua Xu, and Ting-En Lin. Deep open intent classification with adaptive decision boundary. *arXiv preprint arXiv:2012.10209*, 2021.