applied sciences

MDPI

Article

Semantically Guided Enhanced Fusion for Intent Detection and Slot Filling

Songtao Cai [1], Qicheng Ma [1], Yupeng Hou [1] and Guangping Zeng [1,2,*]

[1] School of Computing & Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China; b20180322@xs.ustb.edu.cn (S.C.); d202110388@xs.ustb.edu.cn (Q.M.); d202210392@xs.ustb.edu.cn (Y.H.)

[2] Beijing Key Laboratory of Knowledge Engineering for Materials Science, University of Science and Technology Beijing, Beijing 100083, China

* Correspondence: zgp@ustb.edu.cn

Abstract: Intention detection and slot filling are two major subtasks in building a spoken language understanding (SLU) system. These two tasks are closely related to each other, and information from one will influence the other, establishing a bidirectional contributory relationship. Existing studies have typically modeled the two-way connection between these two tasks simultaneously in a unified framework. However, these studies have merely contributed to the research direction of fully using the correlations between feature information of the two tasks, without sufficient focusing on and utilizing native textual semantics. In this article, we propose a semantic guidance (SG) framework, enabling enhancing the understanding of textual semantics by dynamically gating the information from both tasks to acquire semantic features, ultimately leading to higher joint task accuracy. Experimental results on two widely used public datasets show that our model achieves state-of-the-art performance.

Keywords: spoken language understanding; natural language understanding; intent detection; slot filling

check for updates

Citation: Cai, S.; Ma, Q.; Hou, Y.; Zeng, G. Semantically Guided Enhanced Fusion for Intent Detection and Slot Filling. *Appl. Sci.* **2023**, *13*, 12202. https://doi.org/10.3390/app132212202

Academic Editors: Fei Xiong, Hongshu Chen, Xuzhen Zhu and Shirui Pan

Received: 21 October 2023
Revised: 4 November 2023
Accepted: 8 November 2023
Published: 10 November 2023

cc BY

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Intent detection and slot filling (IDSF) are quintessential components in spoken language understanding (SLU) systems and are instrumental in decoding user inquiries [1]. These tasks are pivotal for natural language understanding (NLU) in man-machine conversational interfaces. Intent detection is tasked with discerning the user's objective from their input, while slot filling involves identifying specific entities that provide detailed context for the intent. For instance, in command "Aircraft No. 3 takeoff to intercept enemy aircraft in the southwestern", the task of intent detection is to identify its intent label of "interception" from a predefined label set. Concurrently, the task of slot filling is to generate slot labels for each word in command, such as "B-our attacking unit", "I-our attacking unit", "E-our attacking unit", "O", and "S-direction". Since errors of IDSF will propagate to downstream tasks such as dialogue state tracking, and ultimately negatively affect the user experience of SLU systems, it is worth pursuing solutions with higher accuracy for IDSF.

In recent years, due to the significant correlation between intent and slot information, numerous studies have made efforts to leverage this connection. Some research exploits intent information as supplementary data to enhance slot filling [2–4], while other works employ joint models to extract slots and intents simultaneously [5–8]. For instance, in the former type of methods, Goo et al. [2] achieved superior semantic framing results by leveraging the relationship between slot-gate learning intentions and slot attention vectors. Li et al. [3] utilized intention-enhanced embeddings, obtained via a neural network with a self-attentive mechanism, as an entry point for marking slot labels. Qin et al. [4] captured intent semantic knowledge by directly employing intent information as input for slot filling. In the latter approach, for instance, Qin et al. [5] modeled the strong correlation between slots and intents by introducing an intent–slot graph interaction layer. Liu et al. [6] incorporated attention to intent and slot information into alignment-based recurrent neural network (RNN) models, furnishing additional data for intent classification and slot label prediction. Xing et al. [7] effectively represented the relationship between semantic nodes and labeled nodes by initializing the labels of two tasks and constructing two heterogeneous graph attention networks over them. Qin et al. [8] considered cross-influence by establishing a two-way connection between the two related tasks, enabling slots and intents to focus on mutual information. However, the aforementioned studies have either focused solely on a unidirectional flow of information from intent to slot, or concentrated on the interplay between these two tasks, thereby overlooking the semantics of the text.

Inspired by He et al.'s [9] introducing knowledge base as complementary for IDSF, we argue that the semantics of text should be adequately utilized to achieve better performance. Specifically, we introduce a semantic guidance (SG) framework for IDSF tasks. To effectively extract slot and intent information, a feature extractor is employed to distill slot and intent information from text into word embedding features. Subsequently, a feature fusion module with a dynamic gating strategy is devised to amalgamate slot information, intent information, and text information. The slot and intent information are harnessed to guide the textual representation, enhancing the accuracy of IDSF. To further validate our proposed method, experiments were conducted on two public datasets. By concurrently investigating inter-modal and intra-modal relationships, it is concluded that the method of utilizing slot information and intention information to guide textual representations significantly augments the performance of the model. The detailed source code of our work is available at https://github.com/USTBSCCE1028/SG (accessed on 20 October 2023).

2. Related Work

In natural language understanding tasks of IDSF, modeling the closely correlated relationship between the two tasks has been a key focus in many state-of-the-art joint models.

These joint modeling approaches can be broadly classified into two categories. The first category involves the unidirectional utilization of intent labels as supplementary information to enhance the effectiveness of slot filling tasks. For instance, Liu and Lane [6] employed an intent-augmented gate mechanism for the slot filling task. Zhang and Wang [10] pioneered the application of RNN in the task of intent detection. Goo et al. [2] employed intent labels to guide the slot gate. Li et al. [3] leveraged intent semantic representations as gate markers for slot labels, resulting in improved performance with the ATIS dataset. Qin et al. [4] employed the stack propagation joint model, which allows for the utilization of intent labels as input for the slot filling task. Zhao et al. [11] employed joint models in the domain of named entity recognition within the medical field. Zhang et al. [12] utilized a unidirectional joint model in downstream tasks of coreference resolution. Ni et al. [13] investigated how to combine specialized models built independently for each task to achieve a complete joint task. They compared the effectiveness of hybrid-based and recurrent convolutional neural network (RCNN) models.

The second category involves using both intent and slot labels bidirectionally to enhance the model's performance on these two correlated tasks. For example, E et al. [14] designed an iterative mechanism for joint tasks in their proposed network. Liu et al. [15] used stacked blocks to achieve a joint global representation of slot and intent information and released a Chinese dataset. Wu et al. [16] introduced a non-autoregressive model to enhance inference speed. Zhang et al. [17] proposed a graph neural network-based model to realize semantic interaction between entities, sentences, and documents. Qin et al. [18] proposed a non-autoregressive model with graph interaction layers. Qin et al. [8] proposed a co-interactive transformer that considers the cross-impact between the two tasks. Xing and Tsang [7] proposed a two-stage network, in the first stage of which initial estimation labels for two tasks are generated, and in the second stage, they are used to mutually guide each other. Ma et al. [19] decomposed slot filling into two stages of slot proposal and slot classification to address the significant differences in slot positions for the same intent in different utterances. Xing and Tsang [20] constructed a heterogeneous label graph containing two topologies and used the proposed ReLa-Net model to capture label correlations. Song et al. [21] fully utilized the statistical co-occurrence frequency between intent and slots as prior knowledge to enhance the joint task. Abro et al. [22] encoded domain knowledge using regular expression rules to enhance the model's performance when transferring to a new domain with limited training data. Hao et al. [23] introduced an intent embedding matrix to guide slot filling tasks with intent information and applied it to their released agricultural dataset AGIS. He et al. [9] integrated the WordNet knowledge base with a BiLSTM model through an attention mechanism, providing additional semantic information to enhance text comprehension. Dao et al. [24] studied the impact of disfluency detection on downstream tasks of IDSF. Tavares et al. [25] further integrated intent inference and slot filling tasks with dialogue state tracking tasks to obtain more accurate dialogue state inference results. Castellucci et al. [26] introduced a multilingual recurrence-less model for joint tasks. Stoica et al. [27] introduced a capsule network structure for Romanian joint tasks. Dao et al. [24] presented the first public IDSF dataset for Vietnamese. Akbari et al. [28] established a Persian benchmark for joint IDSF based on the ATIS dataset. Firdaus et al. [29] proposed a multilingual multitask model that shares sentence encoders among three languages.

While the above studies have focused on modeling the inter-dependencies between intent and slot tasks, they place relatively less emphasis on fully exploiting the textual semantics. Our proposed framework strengthens the role of semantic features by using intent and slot information to dynamically guide the text representations. The focus is on integrating this semantic guidance into the joint modeling.

3. Methodology

This section presents the details of the SG framework designed for IDSF, as illustrated in Figure 1. The framework incorporates various components and strategies, including a feature extractor, a label attention extractor, feature fusion, dynamic gates, and two separate classifiers for IDSF.

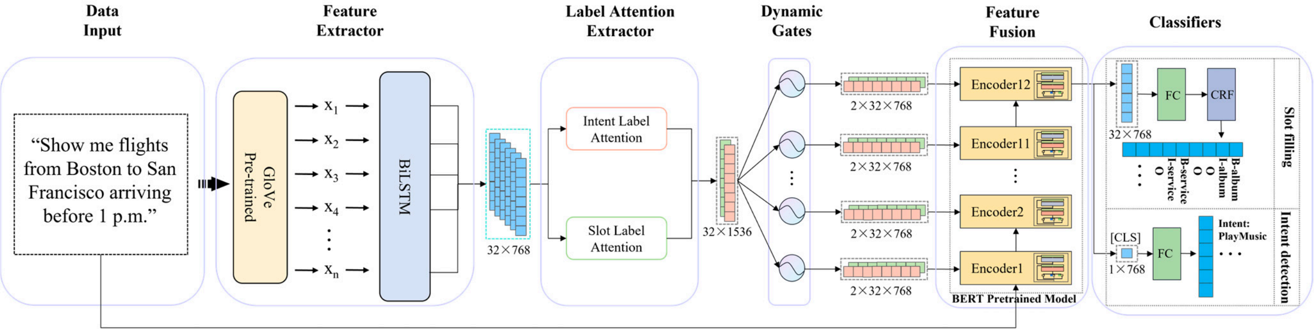

Figure 1. Overview of the proposed semantic guidance framework.

3.1. Compatibility Settings

To ensure the compatibility of each stage in the SG framework, the vector matrices of the different stages are dimensionally transformed. Initially, in the feature extraction stage, the input data are first extracted into a 300-dimensional feature matrix, whose dimension is restricted by the GloVe pre-training model, and later subsequently transformed into a 768-dimensional feature matrix by BiLSTM, thus preparing for the subsequent label attention extraction with the matched shape. The feature matrix is then fed into the label attention extractor, which contains two paralleled label attention blocks with identical input and output shape to the 768, respectively, and specifically for intent and slot label extraction. Subsequently, the two feature matrices are spliced into a 1536-dimensional vector and fed into 12 dynamic gates parallelly. In order to enable the features to be fused as semantic guides for the textual information in BERT, the vector is transformed into two 768-dimensional feature matrices. Finally, in the feature fusion stage, the two 768-dimensional semantic guide vectors, separately acting as key and value, are deeply fused with the same shaped text information as used in BERT.

3.2. Feature Extractor

The 300d GloVe pre-training model [30] is used to extract the initialization vectors of the text, which is a popular word embedding learning method that combines global statistical information and local contextual information to learn word vectors. Subsequently, initialization vectors are fed into BiLSTM to obtain rich contextual information. BiLSTM reads the input sequence forwards and backwards to obtain context-sensitive hidden states at each time step. Specifically, for the input sequence $\{x_1, x_2, x_3, \ldots, x_n\}$ where n is the number of words in the sequence, BiLSTM will generate a series of hidden states $H = \{h_1, h_2, h_3, \ldots, h_n\}$.

3.3. Label Attention Extractor

Building upon the achievements of Cui [31] and Qin [8] et al. in capturing labeled word vector features, we introduce an attention mechanism to create a labeled attention extractor. This extractor enhances the focus on both intention and slot representations. Specifically, the two fully connected layer weight matrices are initialized as weight matrices for intentional and slot filling representations, respectively, as follows:

$$W_s \in M^{d \times n^S} \tag{1}$$

$$W_I \in M^{d \times n^I} \tag{2}$$

where W_s denotes the weight matrix of the slot filling representation, W_I denotes the weight matrix of the intent representation, M denotes matrix, n^I denotes the number of labels for the intent, n^S denotes the number of labels for the slots, and d denotes the hidden layer dimension.

First, the output vectors of the extractor are initialized through matrix multiplication and then activated via the Leaky ReLU activation function to obtain the label importance distributions of the intent and slot representations, the formulas of which are as follows:

$$IntentScore = f(W_I \cdot v_I) \tag{3}$$

$$SlotScore = f(W_S \cdot v_S) \tag{4}$$

where v_I denotes the intent vector features, v_S denotes the slot vector features, $\cdot$ denotes matrix multiplication, and $f(\cdot)$ denotes the activation function of Leaky ReLU, which aims to alleviate the neuron death problem of zero-value gradient by introducing a small positive slope that allows a non-zero output for negative input values. $IntentScore$ and $SlotScore$ denote the labeled importance distribution of intents and slots, respectively.

Finally, the label importance distributions are weighted with the corresponding original fully connected weight matrices to obtain the final intent representation and slot representation, the formulas of which are as follows:

$$F_I = IntentScore + W_I \tag{5}$$

$$F_S = SlotScore + W_S \tag{6}$$

where F_I denotes the intent representation and F_S denotes the slot representation.

3.4. Dynamic Gates and Feature Fusion

Dynamic gates and feature fusion are designed to integrate intent information, slot information, and text information. The main idea is to utilize intent and slot information to guide textual representation. The specific network structure is shown in Figure 2. The intent representation is fused with the slot representation and eventually represented as keys and values, which are used together with the text semantics to compute the attention weights. Specifically, 12 dynamic gates are specifically designed for fusing intent and slot features. It is worth noting that each of these 12 dynamic gates corresponds one of 12 encoders in the BERT [32] model, which helps to integrate the intent and slot features into the serialized textual representation while dynamically adjusting the importance of the intent and slot features. This complex fusion mechanism is able to efficiently integrate and reconcile text, intent, and slot information throughout the model architecture. Specifically, intent and slot features are spliced into a vector, and then fed into the 12 dynamic gates. The expression for the corresponding feature set vector is as follows:

$$F = C(F_I, F_S) \tag{7}$$

where F denotes the feature after intent and slot fusion and $C(\cdot)$ denotes the concatenation operation that stitches two 768-dimensional vectors into one 1536-dimensional vector.

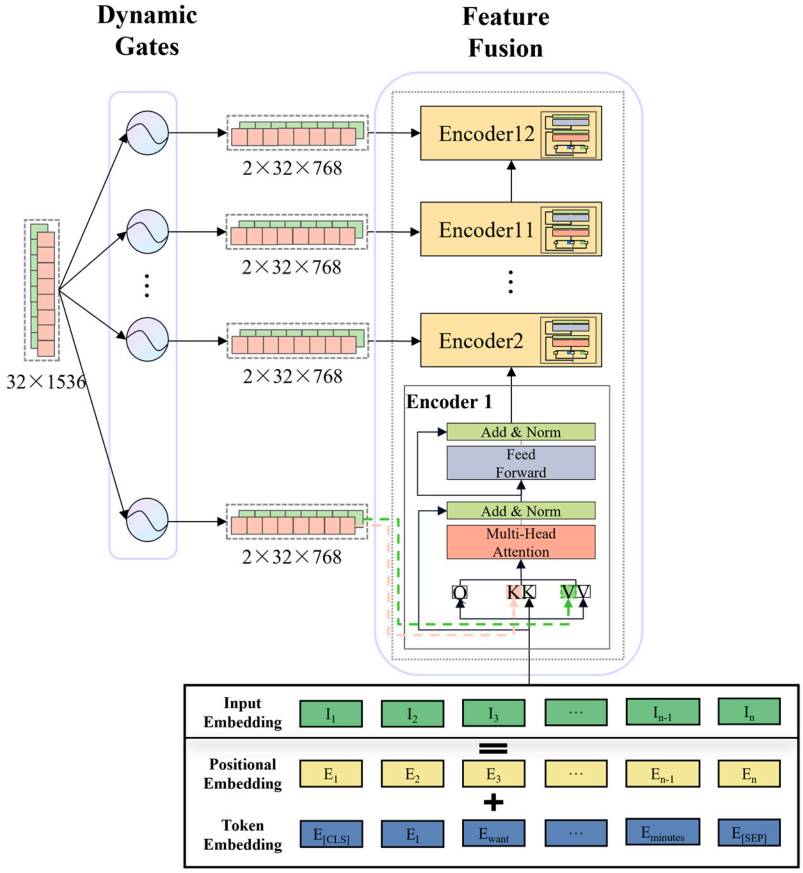

Figure 2. Schematic diagram of the dynamic gates and feature fusion module.

By resetting the weight of the input feature vector, we apply the dynamic gates mechanism to obtain the normalized vector G_i as follows:

$$V_i = F \cdot w_i + b \tag{8}$$

where $V_i \in [0, 1], i \in [1, 12]$ denotes the intent and slot fused features of the output of the i-th dynamic gate, $\mathrm{w}_i \in [0, 1], i \in [1, 12]$ denotes the weight of the i-th dynamic gate corresponding to the i-th layer encoder of BERT, and b denotes bias.

Subsequently, the feature is split into key and value after passing through the activation function, the formulas of which are as follows:

$$Key_i, Value_i = S(f(V_i)) \tag{9}$$

where $f(\cdot)$ denotes the activation function of Leaky ReLU and $S(\cdot)$ denotes the splitting operation, dividing the 1536-dimensional vector output of the dynamic gates unit into two 768-dimensional vectors, which are the input of the BERT framework. Key_i and $Value_i$ denote the key and value of the i-th visual feature corresponding to the i-th encoder of BERT, respectively.

Finally, the contextual query (Q) provided by the textual modality meets the value (V) and key (K) based on the visual modality. Both of them are injected simultaneously into the BERT framework to compute the attention parameters between different modalities.

3.5. Classifiers

The slot filling classifier is specifically designed to accurately predict the class of slots in a given sentence. This task is achieved through a dual approach: the conditional random fields (CRFs) layer and the fully connected (FC) layer. Together, they play a crucial role in identifying diverse slot filling types. The CRFs model is a well-suited choice for sequence annotation tasks, as it aims to assign labels to individual units within a sequence. In our research, the global word vectors generated by BERT are fed into the FC layer to construct a labeled sequence $y = \{y_1, y_2, \ldots, y_n\}$. Then, the CRFs layer calculates the observed conditional probability of a given labeled sequence y based on the hidden vector H_L of BERT. Following maximum likelihood estimation, the prediction result L_{CRF} is shown in the following equation:

$$p(y|H_L) = \frac{\prod_{i=1}^{n} T(y_{i-1}, y_i, H_L)}{\sum_{y' \in Y} \prod_{i=1}^{n} T\left(y'_{i-1}, y'_i, H_L\right)} \tag{10}$$

$$L_{CRF} = -\sum_{i=1}^{M} \log\left(p\left(y^{(i)}|H_L\right)\right) \tag{11}$$

where $T(\cdot)$ denotes the transition function. Given an input sequence H_L and a pair of consecutive labels, for instance, y_{i-1} and y_i, this function computes the percentage probability of the chosen label sequence relative to the entire set of possibilities. Here, Y signifies the predefined set of labels, formulated according to the BIO labeling scheme.

An intent detection classifier is designed to predict the intent type expressed in a sentence. In this paper, the [CLS] vector outputted by BERT is used as sentence vector. It is an approach that has shown significant performance in several natural language processing tasks owing to the richness of semantic information of the vectors [33]. Subsequently, the sentence vector is fed into the FC layer for sentence intent prediction.

4. Experiments

In this section, the experimental results of comparative experiments are shown to demonstrate the effectiveness of the SG framework. In addition, an ablation study is performed for each of our newly proposed architectures.

4.1. Dataset and Settings

In the experimental setup, two benchmark datasets are utilized to evaluate the performance of IDSF tasks, which are ATIS [2] and SNIPS [34]. The ATIS dataset contains recordings of booking flights, whose training set contains 4478 utterances, development set contains 500 utterances, and test set contains 893 utterances. SNIPS was collected from the

Snips personal voice assistant, whose training set contains 13,084 utterances, development set contains 700 utterances, and test set contains 700 utterances.

Some hyperparameters are set initially as follows: the Adam optimizer with a learning rate of 1×10^{-5}, batch size of 32, training epochs of 30, and a cap on sentence length at 32 tokens. Moreover, to convert vector values to probabilities effectively and enhance computational speed, the ReLU activation function is integrated.

4.2. Comparative Experiments

To scientifically evaluate the performance, several baseline models were selected for comparative experiments and compared with our model. The baseline models are listed below.

Slot-Gated [2] focuses on learning the relationship between intents and slots through slot gates, leading to better IDSF task performance in global optimization.

SF-ID Network [14] guarantees a bi-directional correlation between intents and slots through the SF sub-network and ID sub-network. By using an innovative iterative mechanism, the model mutually promotes these two tasks during training, leading to better global optimization.

CM-Net [15] utilizes a novel collaborative memory network that first obtains slot-specific and intent-specific features from memory collaboratively and then uses these features for modeling. By leveraging the co-occurrence relationship between slots and intents, it achieves better global optimization when training end-to-end models.

Stack-Propagation Framework [4] proposes a framework that incorporates token-level intent detection to capture intent semantic knowledge by directly using the intent as input for slot filling. This approach better integrates intent information through a joint model of stack propagation and effectively reduces error propagation, thus improving overall model performance.

Co-Interactive Transformer [8] introduces a Co-Interactive module that establishes a two-way connection between IDSF. Through a specific mutual attention mechanism, intents and slots can pay attention to each other's information, thus enabling deep interaction between intents and slots and improving overall task performance.

The results of the comparison experiments are shown in Table 1.

Table 1. Performance of comparative experiments running on our same local machine.

Model	**ATIS**			**SNIPS**		
	Slot (F_1)	**Intent (Acc)**	**Overall (Acc)**	**Slot (F_1)**	**Intent (Acc)**	**Overall (Acc)**
Slot-Gated * [2]	88.72	95.22	74.80	94.20	93.66	82.90
SF-ID Network * [14]	90.55	96.15	78.84	94.72	96.58	86.24
CM-Net * [15]	93.40	96.75	84.53	95.14	96.24	85.47
Stack-Propagation * [4]	94.25	97.05	86.95	95.44	97.15	86.65
Co-Interactive Transformer * [8]	95.65	96.81	86.45	95.20	98.11	88.92
SG framework *	96.55	97.54	88.20	95.60	98.43	89.30

All asterisks (*) denote results run on the local machine.

Overall, the SG framework model significantly outperforms the benchmark model. On the ATIS dataset, the SG framework improves the F_1 score on the slot filling task by about 1% compared to the other models. Similarly, it improves the correctness of intent detection by about 1%, and improves the overall correctness by about 2%. On the SNIPS dataset, the SG framework improves the F_1 score on the slot filling task by about 1.5% compared to the other models. Similarly, the correctness of intent detection improves by about 1%, and the overall correctness impressively improves by about 1%. The results show that the SG framework achieves better results on both the IDSF tasks. Most importantly, the design of the 12 dynamic gates ensures that the intent and slot features can effectively interact with each encoder layer of BERT to capture richer contextual information. In this way, our framework not only takes advantage of the powerful representational capabilities of BERT, but also ensures that these representational capabilities are fully utilized throughout the model. Dynamic gates can automatically adjust the weights based on the input features, thus realizing adaptive feature fusion. Compared to simple feature splicing or weighted averaging, the dynamic gate mechanism can explore the interactions between different features in greater depth, which not only enhances the expressiveness of the model, but also ensures that critical information is kept in the fusion process.

4.3. Ablation Study

This section presents a series of ablation experiments conducted to evaluate the individual contributions of each component for the overall performance of our proposed model. Specifically, these experiments aim to validate the effectiveness of various modules in the following configurations:

Model 1: Intent label attention is used only as a guide for text semantics. With this configuration, it is possible to compare it with a version that employs both intent label attention and slot label attention strategies.

Model 2: Slot label attention is used only as a guide for text semantics. With this configuration, it is possible to compare it with a version that employs both intent label attention and slot label attention strategies.

Model 3: Dynamic gates are replaced with 12 MLP layers, corresponding to the 12 encoder layers. This configuration can be compared to the version with the dynamic gates strategy.

According to the results in Table 2, for both the ATIS and SNIPS datasets, we can draw the following conclusions:

Table 2. Ablation experiments for the proposed model.

Model	ATIS			SNIPS		
	Slot (F_1)	Intent (Acc)	Overall (Acc)	Slot (F_1)	Intent (Acc)	Overall (Acc)
Model 1	96.14	96.88	87.25	95.33	97.95	88.75
Model 2	95.87	96.35	87.03	95.12	97.50	88.62
Model 3	94.85	95.56	86.32	94.35	97.02	88.21
SG framework	96.55	97.54	88.20	95.60	98.43	89.30

To begin with, using only intent label attention as textual semantic guidance reduces the slot filling F_1 scores by about 0.4% and 0.3%, respectively, the correct rates of intent detection by about 0.7% and 0.5%, respectively, and the overall correct rates by about 1% and 0.5%, respectively. Subsequently, using only slot label attention as text semantic guidance reduces the slot filling F_1 scores by about 0.8% and 0.5%, respectively, the correct rate of intent detection by about 1.2% and 1%, respectively, and the overall correct rate by about 1.2% and 0.7%, respectively. It can be seen that the intent label plays a more crucial role in semantic guidance. The intent label provides a more macroscopic semantic framework for the model, which provides context for specific information in the text, thus helping the model to better interpret this information, whereas the slot label, although useful in some contexts, may rely more on the overall intent information to provide it with semantic context. Notably, the model works best when intent label attention and slot label attention are used jointly. This suggests that while intent labeling alone may be more critical than slot labeling, when the two are combined, they complement each other in capturing the details and overall semantics of the text for optimal performance. Finally, replacing the dynamic gate model with the MLP layer results in a severe performance degradation. Specifically, the F_1 score for slot filling drops by about 1.7% and 1.25%, respectively, the correctness of intent detection drops by about 2% and 1.4%, respectively, and the overall correctness drops by about 2% and 1%, respectively. The experiments demonstrate that this substitution leads to the most significant performance drop among all ablation experiments. This highlights the pivotal function of dynamic gates in influencing model performance. Dynamic gates possess the capability to automatically fine-tune feature weights based on inputs, facilitating adaptive feature fusion, while the MLP layer in isolation lacks this adaptive capability.

4.4. Discussion

The SG framework demonstrates exceptional performance on the ATIS and SNIPS datasets, highlighting its pivotal role in advancing SLU systems. It surpasses traditional models in terms of accuracy for both IDSF tasks, thanks to its dynamic semantic processing capabilities.

With the ATIS dataset, tailored for travel-related inquiries, the SG framework distinguishes between different user intents with finesse. For example, it adeptly discerns a general flight query from an explicit booking request, streamlining responses for users seeking flight information or making travel arrangements. This is crucial for enhancing the efficacy of online booking services.

When applied to the SNIPS dataset, the framework's versatility becomes evident as it accurately processes a wide array of voice commands. For instance, a user's command to play music, like "Play the latest album by Coldplay on Spotify", is interpreted with precision, catering to specific requests and enhancing user interaction within smart home environments.

The adaptability of the SG framework in these scenarios underscores its practical application and potential for integration into various forms of SLU technology. Recognizing the significance of context and dialect variations, future developments will focus on further refining the model's ability to handle the intricacies of natural language. This paves the way for SLU systems that are more intuitive and efficient in terms of user communication across diverse settings.

5. Conclusions

Our proposed SG framework effectively combines intent labeling attention and slot labeling attention to better capture the overall semantics and details of text. Additionally, the dynamic gate strategy is designed to adaptively adjust the weights according to the input features, which in turn enables adaptive feature fusion. Experimental results show that the SG framework significantly outperforms existing state-of-the-art models, achieving about 1% and 1.5% improvements in F_1 scores, respectively. These results demonstrate the effectiveness of our approach and prove the feasibility of enhancing the input semantic information by using intents and slot labels. In future work, we plan to evaluate the performance of the SG framework in different domains and optimize the training speed of the model.

Author Contributions: Conceptualization, S.C.; methodology, S.C. and Q.M.; software, S.C. and Y.H.; validation, S.C., Y.H. and Q.M.; formal analysis, S.C.; investigation, G.Z.; resources, S.C.; data curation, Y.H. and Q.M.; writing—original draft preparation, S.C., Q.M. and Y.H.; writing—review and editing, G.Z.; visualization, Q.M. and Y.H.; supervision, G.Z.; project administration, G.Z.; funding acquisition, G.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (General Program), grant number 62072031, and the Second Department, Second Institute of China Aerospace Science and Industry Corporation, grant number classified.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/USTBSCCE1028/SG (accessed on 20 October 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tur, G.; Mori, R.D. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*; John Wiley & Sons: Hoboken, NJ, USA, 2011; ISBN 978-1-119-99394-0.
2. Goo, C.-W.; Gao, G.; Hsu, Y.-K.; Huo, C.-L.; Chen, T.-C.; Hsu, K.-W.; Chen, Y.-N. 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, Volume 2 (Short Papers), New Orleans, LA, USA, 1–6 June 2018; Association for Computational Linguistics: Kerrville, TX, USA, 2018; pp. 753–757.
3. Li, C.; Li, L.; Qi, J. A Self-Attentive Model with Gate Mechanism for Spoken Language Understanding. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; Association for Computational Linguistics: Kerrville, TX, USA, 2018; pp. 3824–3833.
4. Qin, L.; Che, W.; Li, Y.; Wen, H.; Liu, T. A Stack-Propagation Framework with Token-Level Intent Detection for Spoken Language Understanding. *arXiv* **2019**, arXiv:1909.02188.
5. Qin, L.; Xu, X.; Che, W.; Liu, T. AGIF: An Adaptive Graph-Interactive Framework for Joint Multiple Intent Detection and Slot Filling. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November 2020; Association for Computational Linguistics: Kerrville, TX, USA, 2020; pp. 1807–1816.
6. Liu, B.; Lane, I. Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling. *arXiv* **2016**, arXiv:1609.01454.
7. Xing, B.; Tsang, I.W. Co-Guiding Net: Achieving Mutual Guidances between Multiple Intent Detection and Slot Filling via Heterogeneous Semantics-Label Graphs. *arXiv* **2022**, arXiv:2210.10375.
8. Qin, L.; Liu, T.; Che, W.; Kang, B.; Zhao, S.; Liu, T. A Co-Interactive Transformer for Joint Slot Filling and Intent Detection. In Proceedings of the ICASSP 2021—IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021.
9. He, T.; Xu, X.; Wu, Y.; Wang, H.; Chen, J. Multitask Learning with Knowledge Base for Joint Intent Detection and Slot Filling. *Appl. Sci.* **2021**, *11*, 4887. [CrossRef]
10. Zhang, X.; Wang, H. A Joint Model of Intent Determination and Slot Filling for Spoken Language Understanding. *IJCAI* **2016**, *16*, 2993–2999.
11. Zhao, S.; Liu, T.; Zhao, S.; Wang, F. A Neural Multi-Task Learning Framework to Jointly Model Medical Named Entity Recognition and Normalization. *AAAI* **2019**, *33*, 817–824. [CrossRef]
12. Zhang, R.; dos Santos, C.N.; Yasunaga, M.; Xiang, B.; Radev, D. Neural Coreference Resolution with Deep Biaffine Attention by Joint Mention Detection and Mention Clustering. *arXiv* **2018**, arXiv:1805.04893.
13. Ni, P.; Li, Y.; Li, G.; Chang, V. Natural Language Understanding Approaches Based on Joint Task of Intent Detection and Slot Filling for IoT Voice Interaction. *Neural Comput. Appl.* **2020**, *32*, 16149–16166. [CrossRef]
14. Niu, P.; Chen, Z.; Song, M. A Novel Bi-Directional Interrelated Model for Joint Intent Detection and Slot Filling. *arXiv* **2019**, arXiv:1907.00390.
15. Liu, Y.; Meng, F.; Zhang, J.; Zhou, J.; Chen, Y.; Xu, J. CM-Net: A Novel Collaborative Memory Network for Spoken Language Understanding. *arXiv* **2019**, arXiv:1909.06937.
16. Wu, D.; Ding, L.; Lu, F.; Xie, J. SlotRefine: A Fast Non-Autoregressive Model for Joint Intent Detection and Slot Filling. *arXiv* **2020**, arXiv:2010.02693.
17. Zhang, Q.; Chen, H.; Cai, Y.; Dong, W.; Liu, P. Modeling Graph Neural Networks and Dynamic Role Sorting for Argument Extraction in Documents. *Appl. Sci.* **2023**, *13*, 9257. [CrossRef]
18. Qin, L.; Wei, F.; Xie, T.; Xu, X.; Che, W.; Liu, T. GL-GIN: Fast and Accurate Non-Autoregressive Model for Joint Multiple Intent Detection and Slot Filling. *arXiv* **2021**, arXiv:2106.01925.
19. Ma, Z.; Sun, B.; Li, S. A Two-Stage Selective Fusion Framework for Joint Intent Detection and Slot Filling. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, 1–12. [CrossRef] [PubMed]
20. Xing, B.; Tsang, I.W. Group Is Better than Individual: Exploiting Label Topologies and Label Relations for Joint Multiple Intent Detection and Slot Filling. *arXiv* **2022**, arXiv:2210.10369.
21. Song, M.; Yu, B.; Quangang, L.; Yubin, W.; Liu, T.; Xu, H. Enhancing Joint Multiple Intent Detection and Slot Filling with Global Intent-Slot Co-Occurrence. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 7967–7977.
22. Abro, W.A.; Qi, G.; Aamir, M.; Ali, Z. Joint Intent Detection and Slot Filling Using Weighted Finite State Transducer and BERT. *Appl. Intell.* **2022**, *52*, 17356–17370. [CrossRef]
23. Hao, X.; Wang, L.; Zhu, H.; Guo, X. Joint Agricultural Intent Detection and Slot Filling Based on Enhanced Heterogeneous Attention Mechanism. *Comput. Electron. Agric.* **2023**, *207*, 107756. [CrossRef]
24. Dao, M.H.; Truong, T.H.; Nguyen, D.Q. From Disfluency Detection to Intent Detection and Slot Filling. In Proceedings of the Interspeech 2022, Incheon, Republic of Korea, 18–22 September 2022; pp. 1106–1110.
25. Tavares, D.; Azevedo, P.; Semedo, D.; Sousa, R.; Magalhães, J. Task Conditioned BERT for Joint Intent Detection and Slot-Filling. *arXiv* **2023**, arXiv:2308.06165.
26. Castellucci, G.; Bellomaria, V.; Favalli, A.; Romagnoli, R. Multi-Lingual Intent Detection and Slot Filling in a Joint BERT-Based Model. *arXiv* **2019**, arXiv:1907.02884.

27. Stoica, A.; Kadar, T.; Lemnaru, C.; Potolea, R.; Dînşoreanu, M. Intent Detection and Slot Filling with Capsule Net Architectures for a Romanian Home Assistant. *Sensors* **2021**, *21*, 1230. [CrossRef]
28. Akbari, M.; Karimi, A.H.; Saeedi, T.; Saeidi, Z.; Ghezelbash, K.; Shamsezat, F.; Akbari, M.; Mohades, A. A Persian Benchmark for Joint Intent Detection and Slot Filling. *arXiv* **2023**, arXiv:2303.00408.
29. Firdaus, M.; Ekbal, A.; Cambria, E. Multitask Learning for Multilingual Intent Detection and Slot Filling in Dialogue Systems. *Inf. Fusion* **2023**, *91*, 299–315. [CrossRef]
30. Pennington, J.; Socher, R.; Manning, C. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–19 October 2014; Association for Computational Linguistics: Kerrville, TX, USA, 2014; pp. 1532–1543.
31. Cui, L.; Zhang, Y. Hierarchically-Refined Label Attention Network for Sequence Labeling. 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, 3–7 November 2019; pp. 4113–4126.
32. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
33. Choi, H.; Kim, J.; Joe, S.; Gwon, Y. Evaluation of BERT and ALBERT Sentence Embedding Performance on Downstream NLP Tasks. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10 January 2021; pp. 5482–5487.
34. Coucke, A.; Saade, A.; Ball, A.; Bluche, T.; Caulier, A.; Leroy, D.; Doumouro, C.; Gisselbrecht, T.; Caltagirone, F.; Lavril, T.; et al. Snips Voice Platform: An Embedded Spoken Language Understanding System for Private-by-Design Voice Interfaces. *arXiv* **2018**, arXiv:1805.10190.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.