2023 International Conference on Advanced Computing Technologies and Applications (ICACTA)

Leveraging BERT for Next-Generation Spoken Language Understanding with Joint Intent Classification and Slot Filling

[1]Santosh Gore
Director, Sai Info Solution
Nashik, Maharashtra, India
https://orcid.org/0000-0003-1814-59131
sai.info2009@gmail.com

[2]Dr Devyani Jadhav
Department of Information Technology
Sanjivani College of Engineering
Kopargaon
bhamaredevyaniit@sanjivani.org.in

[3]Mayur Eknath Ingale
Sandip Institute of Technology and Research Centre
Nashik
mayur.ingale@sitrc.org

[4]Sujata Gore
Director
Sai Info Solution, Nashik
Maharashtra, India
sujatarpatil21@gmail.com

[5]Prof. Umesh Nanavare
Department of CSE
School of Computing
MIT ADT University, Pune
umeshubn3@gmail.com

2023 International Conference on Advanced Computing Technologies and Applications (ICACTA) | 979-8-3503-4834-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICACTA58201.2023.10393437

***Abstract*— This paper introduces a new technique to natural language processing using BERT (Bidirectional Encoder Representations from Transformers). The program encodes the input text with BERT to generate contextual representations of words, which can then be utilized for intent classification and slot filling through joint training. With limited labeled training data, rare words may not have enough examples to learn from. The proposed model was evaluated against existing models based on attention-based recurrent neural networks and slot-gated models, and was found to be more accurate in terms of intent classification accuracy, slot filling F1 metric score. The level of accuracy in determining the semantic meaning of an entire sentence, as it relates to the context in which it is used, was evaluated on various benchmark datasets. This technique has been employed in various NLP applications, making it a promising area for further study.**

Keywords*— *BERT, intent classification, slot filling, natural language understanding, deep learning, bidirectional representations, NLP applications

I. INTRODUCTION

Natural language understanding (NLU) is essential for the success of objective-driven spoken dialogue systems, such as smart speakers, as it enables users to accomplish tasks through voice interactions. NLU typically involves intent to create a semantic parse for user utterances, classification and slot filling activities are used.

Intent classification is the task of predicting the intent of a given utterance from a predefined set of intents. It is used to understand the user's intent in natural language processing (NLP). It is a type of text classification that is used to classify a text into one of predefined intents. The goal of intent classification is to identify the user's intent so that the system can provide an appropriate response.

Slot filling is a NLP technique used in building voice assistants and conversational agents. It involves the use of pre-defined labeled slots to capture user inputs and store them as semantic variables. The purpose of slot filling is to allow a conversational agent to understand user queries and provide the most appropriate response.

Intent classification is the task of identifying the intention or goal behind a user's input or query, such as "book a flight" or "find a restaurant." Slot filling involves extracting specific pieces of information from the user's input, such as the destination city or the type of cuisine desired.

Example of Intent classification:

User: Hi! Where can I find a restaurant near me?

Bot: Sure, I can help you with that. What type of cuisine are you looking for?

Example of a Slot filling:

User: I am looking for a hotel

Bot: What city are you looking for a hotel in?

This paper discusses the tasks of intent classification and slot filling in NLU and their interrelationship. Traditional independent models for these tasks suffer from error propagation as they do not consider the mutual relationship between them. The paper introduces a bi-directional joint model for intent classification and slot filling that contains a multi-stage hierarchical process using BERT and a bi-directional NLU mechanism. By utilizing the intent2slot and slot2intent models, the suggested approach accomplishes complete combined benefits for joint intent categorization and slot filling, and is trained on the concurrent loss of slot labeling and intent categorization. The contributions of the paper include a bi-directional joint NLU mechanism and a multi-stage hierarchical process integrating Transformer-based bi-directional NLU mechanism and modeling.

II. RELATED WORK

There has been significant prior work on language models with prior training for NLP tasks, particularly with the release of the BERT model. Some of the most influential pre-trained models include ELMo [1] and Generative Pre-trained Transformer (GPT) [2]. ELMo was the first pre-trained language model to use bidirectional LSTMs, and it was demonstrated to produce cutting-edge outcomes on a variety of NLP tasks. GPT, on the other hand, uses a generative language modeling objective to pre-train a large-scale transformer-based language model.

Redford et. al. [3] shows how language models trained on a diverse dataset called WebText can perform natural language processing tasks without explicit supervision. The

largest model, GPT-2, achieves state-of-the-art results on 7 out of 8 language modeling datasets in a zero-shot setting.

Jun et. al. [4] evaluates the ANNA pre-trained language representation model's performance on NLP tasks using a neighbor-aware mechanism to capture context. ANNA outperforms other pre-trained language models on the SQuAD 1.1 and SQuAD 2.0 benchmarks for question answering. Further research is needed to confirm ANNA's competitiveness for other NLP tasks and its robustness for real-world business question answering tasks.

Gunaratna et. al. [5] proposes a new approach for joint intent detection and slot filling in NLU that improves accuracy and provides fine-grained explanations. It uses a collection of binary classifiers for slot type-specific feature learning and is evaluated on two datasets. The model is inherently explainable and the first to explain how slots are filled decisions without post-hoc processing. Extending the model is a task for the future to explain intent detection and exploring its use in other tasks like text classification and named entity recognition (NER).

Hakkani-Tür et. al. [6] proposes a bi-directional RNN-LSTM architecture to estimate complete semantic frames of user utterances in a conversational system. It contributes by investigating alternative architectures for modeling lexical context and providing a joint multi-domain model that enables multi-task deep learning. The proposed architecture shows better performance than other methods on Microsoft Cortana real user data. The approach has the potential to handle belief state updates and non-lexical contexts in one holistic model, leading to an end-to-end conversational understanding framework.

Liu and Lane [7] presents a novel attention-based neural network model for joint intent detection and slot filling, using the ATIS task dataset. The model utilizes an encoder-decoder framework with attention mechanism and alignment-based RNN models to achieve the best-in-class performance for both tasks. The proposed joint training model shows an absolute error decrease on intent detection of 0.56% and an absolute gain of 0.23% on slot filling compared to independent training models. This research proposes a method to add alignment information to attention-based neural network models to improve intent detection and slot filling, enabling better speech and dialog understanding.

Goo et. al. [8] proposes slot-gated mechanism to improve semantic frame results, outperforms baseline on ATIS and SNIPS datasets by 4.2% and 1.9% respectively. Allows for explicit learning of intent-slot relations, filling an important gap in SLU work. An innovative neural network model for joint slot filling and intent detection that preserves the orderly relationship between words, slots, and intents proposed by

Zhang et. al. [9]. It uses dynamic routing-by-agreement and a rerouting schema to improve performance and outperforms existing models. It fills a gap in current efforts that handle intent detection and slot filling separately or do not explicitly preserve the semantic hierarchy.

Chen et. al. [10] proposes a simultaneous intent classification and slot filling model based on BERT to enhance the generalization potential of NLU models. The proposed model outperforms existing models on public datasets and provides guidance for future research on using external knowledge with BERT for more complex NLU datasets.

Qin et. al. [11] proposes a framework for spoken language understanding (SLU) utilizing a combined model with token-level intent detection and StackPropagation. The suggested model performs at the cutting edge on two publicly accessible datasets and incorporates the BERT model. The research gap is to explore the effect of stronger pretrained models in SLU tasks.

III. METHODOLOGY

A. *Input layer that maps words to numerical representations.*

The methodology comprises of a two-way joint model for classifying intent and slot filling. The user's utterances as text phrases in the input data, is tokenized and converted to numerical representations using a pre-trained BERT-BASE model. The suggested approach combines a multi-stage hierarchical process via BERT and bi-directional joint natural language understanding techniques, such as intent2slot and slot2intent, to achieve mutual performance enhancement between intent categorization and slot filling. The model is fed the feature size N, which includes the BERT [CLS] and [SEP] tokens. On two benchmark datasets, ATIS and SNIPS, the model obtains cutting-edge performance in intent classification accuracy, slot filling F1 score, and sentence-level semantic frame accuracy. The model is implemented using deep contextual embeddings and the transformer architecture. The proposed model could be extended to explain intent detection and explore its use in other NLU tasks.

B. *Bi-Directional NLU Modelling Layer*

On top of the NLU Modelling Layer, which contains two models—intent2slot for slot filling and slot2intent for intent classification—we propose a Bi-Directional NLU layer. The intent2slot model uses the full sequence's semantic information to identify each slot and is based on the probability distribution of intent. The slot2intent incorporates the sequence label probability distributions as additional data for intent categorisation. In order to jointly simulate intent classification and slot filling, the model is trained on the joint loss of slot labeling and intent classification. The suggested model uses slot probabilities and BERT to derive initial intent probability. The particular BERT type features twelve encoder stacks, 12 attention heads, and a 768-dimension final hidden layer. A two-layer feed-forward neural network and a multi-head attention module are both present in each encoder stack. It uses a chance of 0.1 for attention dropout, and the output of the second layer is expanded to include residual connection. The output shape is $[N, \ D_E] \ \in R^{N \times DE}$. To prevent potential overfitting, including a layer with a 0.1 dropout rate.

Intent2Slot Model: The intent2slot model extracts semantic information from a sequence to determine the probability of intent and slot labels. The encoder output is denoted as ($h1, h2, \ldots, h_N$) where each hi represents a token's hidden layer state. To classify the complete sequence, the [CLS] token has been trained, and the intent probability P_I is obtained by passing $h1$ across a thick layer and applying softmax classifier.

$$P_I = \text{softmax}(h_I \times W^T{}_I + b_I) \quad (1)$$

The dense layer has DIP nodes represented by a matrix of weights WI $\in$R DIP $\times$DE and a bias vector bI $\in$R DIP. The resulting vector of probabilities PI $\in$R DIP is broadcasted to the length of sequence to be labeled by repeating it N-1 times to get P^ I.

$$\hat{P}_I = \text{repeat}(P_I), N-1 \text{ times} \quad (2)$$

We recover the hidden state sequence of $N-1$ tokens, each with a dimension of D_E, to develop intent2slot:

$$H = (h_2, h_3, \ldots\ldots, h_N) \quad (3)$$

To obtain the slot probability for each word in intent2slot, we combine each element of the hidden state sequence H with P_I by concatenating them, and then input the result into a softmax classifier.

$$S = (H \oplus \hat{P}I) \bullet V^T_S + m_S \quad (4)$$

To obtain the slot probability for each word in the input sequence, we create a matrix $S \in R^{(N-1)\times D_{SP}}$ with dimensions $(N-1)$ x $[D_E + D_{IP}]$, where the intent probability is concatenated with a BERT encoding in each row. A dense layer of size D_{SP} is applied to each row using an array of $N-1$ matrices of weights V_S with dimensions $R^{(N-1)\text{ x }DSP\text{ x }(DE+DIP)}$, and an array of $N-1$ vectors of biases m_S with dimensions $(N-1)$ x D_{SP}. The $N-1$ linear outputs are then fed into $N-1$ softmaxes, the projected slot label for each relevant element of the input sequence being provided by each argmax.

$$S_n = \text{argmax}(\text{softmax}(S^n)), n \in \{2, \ldots, N\} \quad (5)$$

This means that S_n is a matrix with first axis as S.

Slot2Intent Model: The slot2intent model is a model that incorporates the slot label probability distributions into intent detection. It does so by taking the tokens' last hidden state sequence, excluding the [CLS] token from a BERT model, where each hidden state has dimension D_E.

$$H = (h2, h3, \ldots, h_N) \quad (6)$$

Each hidden state $h_n \in$ H is passed through a dense layer with DSP nodes and softmax is applied to obtain slot label probability distributions for each token.

$$P_{Sn} = \text{softmax}(h_n \times \mathbb{W}^T_{Sn} + b_{Sn}), n \in \{2, \ldots, N\} \quad (7)$$

This produces a matrix of size $(N-1) \times D_{SP}$ where each row corresponds to a token in the input sequence (excluding the [CLS] token) and each column corresponds to a slot label. These probability distributions are then flattened to obtain a vector of length $\mathbb{W}_S \in D_{SP}$ x (N-1), denoted as $\hat{P}_S$.

$$\hat{P}_I = \text{flatten}(\text{PS}_n), n \in \{2, \ldots\ldots, N\} \quad (8)$$

The intent hidden state is obtained from the [CLS] token's final hidden state and is concatenated with $\hat{P}S$. This concatenated vector is passed through a softmax classifier.

$$I = \text{argmax}(\text{softmax}((h1 \oplus \hat{P}S) \times V^T_I + m_I)) \quad (9)$$

The output is a $\mathbb{V}_I$ vector of length D_{IP} where each entry corresponds to an intent label. The intent label with the highest probability is selected as the predicted intent label.

Joint optimisation: The intent categorization and slot filling tasks are jointly optimized by the slot2intent model. The slot labeling loss is added to train the model. (LS) and the intent classification loss (LI), where L is the total loss and is defined as:

$$L_I = -\sum_{i \in I} I(y_g = i) \log y_i \quad (10)$$

y_i is the output from the softmax classifier in Equation (9). The model learns to minimize this overall loss during training.

$$L_S = -\sum_{n=2}^{N} \sum_{s \in S} I(y^n_g = s) \log y^n_s \quad (11)$$

where $I(y^n_g = s)$ is the indicator function.

The slot loss is calculated by summing over all tokens in the sequence and all possible slot labels. The cross-entropy loss function is used to calculate how much the genuine labels deviate from the projected slot labels.

IV. EXPERIMENTATION DETAILS

The experiment portion uses the ATIS and SNIPS datasets to assess a bi-directional joint model for intent classification and slot filling to cutting-edge models and performs an ablation study. The BERT model used has a dropout rate of 0.1, a weight initialization standard deviation of 0.02, and is trained for 10 epochs on ATIS and 20 epochs on SNIPS. The evaluation metrics used include span-based slot f1-score, intent accuracy, and semantic accuracy, and the Python conlleval script is used with the "seqeval" package to calculate F1-score. Results reveal that the suggested model performs better than the most recent models, and the ablation study shows the contributions of different layers towards performance.

A. Experiment 1: Overall assessment of the joint model

On two datasets, a joint model for intent classification and slot filling was assessed and contrasted with earlier state-of-the-art models. It made use of the suggested bi-directional contextual contribution (slot2intent, intent2slot).

B. Experiment 2: access intent classification

In the intent classification evaluation, baseline models were trained on the SNIPS dataset with 20 epochs and the intent detection F1 score was measured.

C. Experiment 3: replacing certain components

We conducted an ablation study on the ATIS dataset to investigate the effect of different components of the model. They swapped out the input embedding layer with GloVe or word2vec, the NLU modeling layer with LSTM or GRU, and tested each combination with or without the bidirectional NLU layer.

V. RESULTS AND DISCUSSION

A. Result for Experiment 1:

The proposed model outperformed the previous models on all tasks, with improvements ranging from 1.1% to 11.9%. The model also outperformed other models like joint BERT and Stack-Propagation BERT. The results show that Utilizing slots specific to certain intents still enhances the performance of intent classification, even though the ATIS dataset has imbalanced intent label distribution and a high number of mutually shared slots. The longer sequence duration and more slots in ATIS may have contributed to the relatively lesser improvement in the slot filling F1 score. The improvement nevertheless confirms the usefulness of the intent2slot flow.

In comparison to ATIS, SNIPS has a larger vocabulary size due to its cross-domain topics, with relatively higher accuracy of 99.2% for intent classification and a smaller number of overlapping slots (11 vs 79). SNIPS has 72 slots for

7 intents with a maximum slot type count of 15 for BookRestaurant and a minimum of 2 for SearchCreativeWork. The proposed model improves slot filling performance with intent2slot flow, but the improvement is slightly lower than ATIS. Slot value ambiguity in SNIPS (e.g. location entity labeled as CITY, STATE, or COUNTRY) without an explicit dictionary may cause confusion for the machine.

B. *Result for Experiment 2:*

The study found that the F1 scores for intent SearchCreativeWork and SearchScreeningEvent were lower due to non-discriminative slots in their training datasets, leading to confusion and misclassification. Capsule-NN outperformed Slot-gated only for intent RateBook, which has a higher proportion of unique slots. For all intents and purposes, the proposed model produced better F1 scores, which confirms the effectiveness of slot2intent and mutual augmentation.

Tab.1 shows the performance comparison of various models on ATIS and SNIPS datasets for slot filling (measured in F1 score), intent detection (measured in accuracy), and sentence classification (measured in accuracy).

TABLE 1 PERFORMANCE OF NLU ON THE ATIS AND SNIPS DATASETS

Model	ATIS - 10 epoch			SNIPS - 20 epoch		
	Slot (F1)	*Intent (acc)*	*Sent (acc)*	*Slot (F1)*	*Intent (acc)*	*Sent (acc)*
Joint Seq.	94.3	92.6	80.7	87.3	96.9	73.2
Joint Attention	94.2	91.1	78.9	87.8	96.7	74.1
Slot Gated Filling	94.8	93.6	82.2	88.8	97	75.5
Slot Gated Intent	95.2	94.1	82.6	88.3	96.8	74.6
Capsule NN	95.2	95	83.4	91.8	97.3	80.9
Joint BERT	96.1	97.5	88.2	97	98.6	92.8
Stack Propagation	96.1	97.5	88.6	97	99	92.9
Our Model	97.7	98.7	89.3	98.3	99.6	93.4

C. *Result for Experiment 3:*

On both datasets, the combination of "BERT NLU modeling + BERT embedding + Bidirectional NLU layer" outperformed all others, while changing the BERT NLU modeling and BERT embedding to different combinations had a negative impact on performance. This demonstrates how well BERT works for word embedding and NLU modeling. The GloVe and word2vec models in particular saw an improvement in performance overall because to the bidirectional NLU layer. The results of the ablation study highlight the importance of using BERT for word embedding and NLU modeling, as well as the effectiveness of the bidirectional NLU layer.

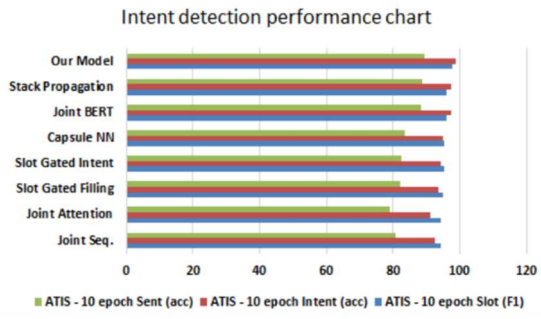

Fig. 1 Performance Comparison chart of Proposed Model

Figure 1 shows a breakdown of intent detection performance of the proposed model and previous state-of-the-art models on the ATIS dataset.

VI. CONCLUSION

The proposed joint model for intent classification can be concluded from the trials and findings discussed and slot filling with bi-directional contextual contribution (slot2intent, intent2slot) superior than current cutting-edge models and other BERT-based joint models. The model was evaluated on two datasets, ATIS and SNIPS, and showed significant improvements in intent classification performance, particularly for SNIPS which has a larger vocabulary size and relatively higher accuracy but also more ambiguity in slot values. The ablation study conducted on the ATIS dataset demonstrated the effectiveness of BERT for word embedding and NLU modeling, as well as the significance of bidirectional NLU layer for overall performance improvement, especially for GloVe and word2vec models. The study also revealed that unique slots still contribute to intent classification performance improvement, and that confusion and misclassification can result from non-discriminative slots in training datasets. Overall, the proposed model and its components have shown promising results when it comes to natural language understanding, particularly in intent classification and slot filling tasks.

VII. SCOPE FOR FURTHER STUDY

While this study has shown that the suggested joint model and bi-directional NLU layer are effective for the tasks of intent classification and slot filling, there is still room for further research. One potential avenue for future work is to investigate the model's performance on more diverse and complex datasets with a wider range of intents and slots, as well as exploring the use of multi-task learning and transfer learning techniques to improve performance on smaller datasets. There is potential to explore different types of embeddings, such as contextualized embeddings, and to incorporate additional sources of knowledge such as ontologies or external knowledge bases. Finally, there is also scope for exploring the use of explainable AI techniques to help understand and interpret the model's decision-making process, particularly in high-stakes applications such as healthcare or finance. This study provides a strong foundation for future research in this area, with the potential to improve the accuracy and robustness of NLU models for a range of applications.

REFERENCES

[1] M. E. Peters *et al.*, "Deep contextualized word representations," *NAACL HLT 2018 - 2018 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 2227–2237, 2018, doi: 10.18653/v1/n18-1202.

[2] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," *Comput. Sci.*, 2018.

[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," *Comput. Sci.*, 2019.

[4] C. Jun *et al.*, "ANNA: Enhanced Language Representation for Question Answering," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 121–132, 2022, doi: 10.18653/v1/2022.repl4nlp-1.13.

[5] K. Gunaratna, V. Srinivasan, A. Yerukola, and H. Jin, "Explainable Slot Type Attentions to Improve Joint Intent Detection and Slot Filling," *Find. Assoc. Comput. Linguist. EMNLP 2022*, pp. 3367–3378, 2022.

[6] D. Hakkani-Tür *et al.*, "Multi-domain joint semantic frame parsing using bi-directional RNN-LSTM," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-September-2016, pp. 715–719, 2016, doi: 10.21437/Interspeech.2016-402.

[7] B. Liu and I. Lane, "Attention-based recurrent neural network models for joint intent detection and slot filling," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-September-2016, no. 1, pp. 685–689, 2016, doi: 10.21437/Interspeech.2016-135.

[8] C. W. Goo *et al.*, "Slot-gated modeling for joint slot filling and intent prediction," *NAACL HLT 2018 - 2018 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 2, pp. 753–757, 2018, doi: 10.18653/v1/n18-2118.

[9] C. Zhang, Y. Li, N. Du, W. Fan, and P. S. Yu, "Joint slot filling and intent detection via capsule neural networks," *ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf.*, pp. 5259–5267, 2020, doi: 10.18653/v1/p19-1519.

[10] Q. Chen, Z. Zhuo, and W. Wang, "BERT for Joint Intent Classification and Slot Filling," 2019, [Online]. Available: arxiv.org/abs/1902.10909

[11] L. Qin, W. Che, Y. Li, H. Wen, and T. Liu, "A stack-propagation framework with token-level intent detection for spoken language understanding," *EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 2078–2087, 2019, doi: 10.18653/v1/d19-1214.