Slot filling and Intent classification - Lab 5

Marco Prosperi (257857)

University of Trento

marco.prosperi@studenti.unitn.it

1. Introduction

The first part of the NLU project involved enhancing the IAS model for intent classification and slot filling on the ATIS dataset [1], evaluated using F1-score and accuracy. Two key features were added: a BiLSTM layer, which processes input sequences bidirectionally to better capture contextual information, and dropout layers to prevent overfitting by randomly deactivating neurons during training. These changes aimed to improve the model's generalization on unseen data. In the second part, the LSTM architecture was replaced with a BERT-based model. This involved fine-tuning and handling sub-tokenization, leveraging BERT's contextual embeddings to boost performance in both tasks.

2. Implementation details

For the first part of the project, I've added incremental changes to the base IAS model. First, I added a bidirectional layer to the LSTM architecture, allowing the model to process the input sequence in both forward and backward directions. This change was implemented by setting the bidirectional parameter to True in the LSTM layer and by multiplying the output by 2 because it concatenates the forward and backward outputs. Dropout layers were also introduced after the embedding and LSTM layers to reduce overfitting and improve generalization.

In the second part, the LSTM architecture was replaced with a BERT-based model, following the joint learning approach described by Chen et al. [2]. The architecture consists of a pre-trained BERT encoder, with two task-specific heads: one for intent classification (using the [CLS] token) and one for slot filling (using the contextualized embeddings of each token). A key challenge was aligning slot labels with BERT's subtokenization (WordPiece). For each word split into sub-tokens, only the first sub-token was assigned the original corresponding slot label, while the others were masked out from the loss computation using the ignore index (-100). This ensured that the slot filling loss was computed only on valid positions.

The implemented model, named *JointBERT*, is based on the `BertPreTrainedModel` class from the Transformers library. It comprises a BERT encoder, followed by two linear layers: one dedicated to intent classification and the other to slot filling. A dropout layer is applied before both classification heads to reduce overfitting. During the forward pass, the model first processes the input through BERT to obtain the last hidden states and the pooled [CLS] output. The pooled [CLS] embedding is fed to the intent classifier, while the full sequence of hidden states is fed to the slot classifier. The final loss is computed as the sum of the intent classification loss and the slot filling loss, where only the active tokens (those corresponding to valid input words) contribute to the slot loss computation.

The model was fine-tuned end-to-end on the ATIS dataset using the AdamW optimizer with a learning rate of 2×10^{-5}, a batch size of 16, and early stopping based on the slot F1-score on the validation set. The best model for each run was saved in a dedicated folder, and evaluation was performed on the official ATIS test set. Additionally, experiments were conducted using both the BERT base and large versions, with the large model achieving superior results at the cost of longer training time.

3. Results

The IAS Model was trained for 200 epoch over 5 runs, with early stopping based on the slot F1-score on the validation set. The learning rate was set to 0.0001, with a batch size of 128 for training and 64 for validation and testing. Results of the first part of the project are shown in Table 1.

Model	Mean Slot F1	Mean Intent Acc.
Baseline IAS	0.925	0.934
IAS + Bidirectionality	0.928	0.936
IAS + Bidirectionality + Dropout	0.94	0.941

Table 1: *Results for Part A: mean slot F1 and intent accuracy for each model variant.*

The BERT-based joint model achieved strong results on the ATIS dataset. Over three runs, for the bert-base-uncased model, the average slot filling F1-score was **0.951**, with a standard deviation of **0.002**, and the average intent classification accuracy was **0.969**, with a standard deviation of **0.006**. The bert-large-uncased model performed even better, achieving an average slot filling F1-score of **0.955** (std: **0.004**) and an average intent classification accuracy of **0.978** (std: **0.004**). These results confirm the effectiveness of BERT for joint NLU tasks, significantly outperforming the LSTM baseline in both metrics.

Model	Slot F1	Intent Acc.
JointBERT (base)	0.951	0.969
JointBERT (large)	0.955	0.978

Table 2: *Comparison of BERT-based models on ATIS.*

For the BERT-based model, the training was performed for 3 epochs with a batch size of 16 for training, 32 for testing and validation, and a learning rate of 2×10^{-5}.

4. References

[1] UniTrento, "Natural language understanding labs," 2025. [Online]. Available: https://github.com/BrownFortress/NLU-2025-Labs/tree/main

[2] Q. Chen, Z. Zhuo, and W. Wang, "Bert for joint intent classification and slot filling," 2019. [Online]. Available: https://arxiv.org/abs/1902.10909