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 implemented incremental improvements to the base IAS model. First, I added bidirectionality to the LSTM architecture by setting the `bidirectional` parameter to True, allowing the model to process input sequences in both forward and backward directions. This required doubling the output dimension of the intent classifier to accommodate the concatenated forward and backward hidden states. Dropout layers (probability 0.1) were introduced after the embedding and LSTM layers to reduce overfitting and improve generalization. The model was trained using the Adam optimizer with a learning rate of 0.0001 and early stopping based on validation slot F1-score, ensuring optimal convergence without overfitting.

In the second part, I replaced the LSTM architecture with a BERT-based model, following the joint learning approach described by Chen et al. [2]. The *JointBERT* model, implemented using the `BertPreTrainedModel` class, consists of a pre-trained BERT encoder with two task-specific classification heads: one for intent classification using the [CLS] token representation, and one for slot filling using the contextualized token embeddings. The architecture leverages shared representations from BERT's transformer layers, enabling effective knowledge transfer between the two tasks.

A critical implementation challenge was handling BERT's WordPiece sub-tokenization for slot filling. The `ATISDataset` class addresses this by using the tokenizer's `word_ids` mapping to assign slot labels only to the first sub-token of each original word, while masking subsequent sub-tokens and special tokens with the ignore index (-100). This ensures that the slot filling loss is computed exclusively on meaningful token positions. The dataset class also handles variable-length sequences through padding and truncation, maintaining consistency across batches.

During training, the model optimizes a joint loss function combining CrossEntropy losses for both tasks. The slot loss is calculated only on active tokens (where `attention_mask == 1` and `slot_labels != -100`), effectively ignoring padded positions and sub-tokens. The training loop implements early stopping based on slot F1-score validation performance, with the best model from each run saved automatically. I conducted experiments with both `bert-base-uncased` and `bert-large-uncased`, using AdamW optimizer with learning rate 2×10^{-5} and batch sizes optimized for each model variant. The training process involved multiple runs to ensure statistical robustness, with each run initialized from different random seeds to account for training variance.

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	0.939	0.943
IAS + Bidirectionality + Dropout	0.942	0.95

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)	0.953	0.972
JointBERT (large)	0.955	0.978

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

For the BERT-based model, the training was performed with different values of batch sizes and the best model was the one trained 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