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

This project addresses the joint learning of intent classification and slot filling through two distinct architectural approaches: enhanced LSTM models and BERT-based transformers, both evaluated on the ATIS dataset.

Part A: LSTM Architecture Evolution. The baseline IAS model underwent systematic enhancement through architectural modifications. First, I replaced the unidirectional LSTM with a bidirectional variant that processes sequences in forward and backward directions simultaneously. This bidirectionality necessitated doubling the intent classifier's input dimension to accommodate concatenated hidden states from both directions. Subsequently, I incorporated dropout regularization (probability 0.1) after embedding and LSTM layers to mitigate overfitting. The training protocol employed Adam optimization with a learning rate of 0.0001 and early stopping based on validation slot F1-score convergence.

Part B: Joint BERT Framework. Drawing inspiration from Chen et al. [2], I implemented a unified transformer architecture that shares contextual representations across both tasks. The *JointBERT* model extends `BertPreTrainedModel` with specialized classification heads: intent prediction utilizes the [CLS] token's pooled representation, while slot filling applies token-level classification to all contextualized embeddings. This design enables efficient joint learning while maintaining task-specific optimization.

Sub-tokenization and Alignment. BERT's WordPiece tokenization creates a fundamental challenge for slot labeling, as word-level annotations must align with sub-token sequences. The `ATISDataset` class addresses this through careful label alignment: using the tokenizer's `word_ids` mapping, labels are assigned exclusively to first sub-tokens of each word, while continuation tokens receive the ignore index (-100). This masking strategy ensures loss computation occurs only on semantically meaningful positions, preserving label integrity during sequence padding and truncation.

Optimization and Robustness. Joint training employs combined CrossEntropy losses for both tasks, computed selectively where attention masks indicate valid tokens and slot labels differ from the ignore index. I systematically evaluated `bert-base-uncased` and `bert-large-uncased` variants using AdamW optimization with learning rates of 2×10^{-5}. Statistical robustness was ensured through multiple training runs with randomized seeds, incorporating early stopping mechanisms to preserve optimal validation performance and prevent overfitting.

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