# NLU Course Projects - NLU

*Ettore Saggiorato (247178)*

University of Trento

ettore.saggiorato@studenti.unitn.it

## 1. Introduction

This assignment focuses on performing slot filling and intention prediction using the ATIS dataset. The objective of this assignment is to implement and evaluate various techniques designed to enhance model performance by bidirectionality and dropout layers to evaluate their impact on performance. Finally fine-tuning a pre-trained BERT model [1] for the same task, this introduces additional complexity managing sub-tokenization.

Model performance is evaluated with accuracy for intent classification and F1 score for slot filling.

## 2. Implementation Details

### 2.1. Part 1 - LSTM Model

In the first part, the impact of architectural enhancements like bidirectionality and dropout are explored. These are applied sequentially to observe their contributions.

#### *2.1.1. Architecture*

**Bidirectionality**'s implementation is straightforward using pytorch's LSTM, but now the model works with double the representations for each token. It's expected to have better performances as the model can not capture richer contextual information.

#### *2.1.2. Regularization*

**Dropout:** Dropout is applied both after the embedding and before the output layers. It's expected to make the model more robust, thus increasing performances.

### 2.2. Part 2 - BERT Fine-tuning

The second part of the assignment involves fine-tuning a pre-trained BERT model [1] (`bert-base-uncased`) for intent prediction and slot filling. Substituting BERT to the LSTM is straightforward. The 'sub-tokenization issue' was managed by setting only the first subtoken of a word as the relevant tag, while all subsequent subtokens were treated as padding (ID = 0).

Training is performed similarly to JointBERT [2]: max 10 epochs, with early stopping, learning rate $5e-5, 3e-5, 2e-5$ and batch size 16[1]. To explore regularization, gradient clipping and learning rate scheduling with 4 epochs of warm-up were applied. All experiments are run 5 times and results are averaged for consistency.

## 3. Results

Table 1 shows the results for the experiments.

### 3.1. LSTM Model

The base model already performs pretty well. Bidirectionality significantly improves performance. Dropout alone doesn't perform as well as bidirectionality, but adding it to 'bi' gets slightly better results overall.

### 3.2. BERT Fine-tuning

The results clearly demonstrate the positive impact of gradient clipping and learning rate scheduling. Training the model it was noted that finetuning with bigger batch-sizes, like 32, performs worse on this task. Results are 'comparable' with JointBERT [3].

## 4. References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[2] P. Jangwon, "Jointbert," https://github.com/monologg/JointBERT/tree/master.

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

| Part | Experiment | LR | Epochs | F1 Slot | Accuracy Intent |
|---|---|---|---|---|---|
| 1.0 | Baseline | 0.0001 | 195 | 92.2 $\pm$0.3 | **93.3** $\pm$0.2 |
| 1.0 | Baseline | 0.001 | 80 | **92.4** $\pm$0.5 | 93.3 $\pm$0.6 |
| 1.0 | Baseline | 0.01 | 40 | 91.8 $\pm$0.4 | 91.6 $\pm$0.2 |
| 1.1 | Bi | 0.001 | 50 | **94.1** $\pm$0.2 | **94.6** $\pm$0.6 |
| 1.2 | DropEmb | 0.001 | 130 | 93.1 $\pm$0.3 | 94.0 $\pm$0.1 |
| 1.2 | DropOut | 0.001 | 80 | 92.8 $\pm$0.5 | 93.5 $\pm$0.4 |
| 1.2 | Drop | 0.001 | 100 | **93.2** $\pm$0.3 | **94.2** $\pm$0.3 |
| 1.3 | BiDrop | 0.001 | 55 | **94.5** $\pm$0.2 | **94.9** $\pm$0.4 |
| 2.0 | BertBase | 5e-2 | 9 | 91.8 $\pm$0.03 | 96.2 $\pm$0.03 |
| 2.0 | BertBase | 5e-3 | 9 | 90.9 $\pm$0.09 | 96.5 $\pm$0.1 |
| 2.0 | BertBase | 5e-5 | 9 | 91.6 $\pm$0.01 | **97.4** $\pm$0.04 |
| 2.1 | BertSch | 5e-5 | 9 | 91.44 $\pm$0.07 | 96.5 $\pm$0.03 |
| 2.2 | BertClip | 5e-5 | 9 | 94.2 $\pm$0.03 | 96.8 $\pm$0.02 |
| 2.3 | BertSchClip | 5e-5 | 9 | **<u>95.2</u>** $\pm$0.01 | <u>97.2</u> $\pm$0.00 |
| 2.3 | BertSchClip32 | 5e-5 | 9 | 93.7 $\pm$0.05 | 97.2 $\pm$0.1 |
| $\star$ | JointBERT | - | - | 96.1 | 97.5 |

Table 1: *Performance of the models for each configuration.* ***Bold*** *values represent the best model for each section,* <u>*underline*</u> *the best overall. Comparison with JointBERT[3] is provided.*