NLU Course Projects - SA

Ettore Saggiorato (247178)

University of Trento

ettore.saggiorato@studenti.unitn.it

1. Introduction

This assignment focuses on fine-tuning of BERT [1] for aspect term extraction in Aspect-Based Sentiment Analysis (ABSA) on the `laptop14` partition of the SemEval-2014 Task 4 dataset [2].

The evaluation of model performance is conducted using standard metrics: precision, recall, and F1 score, all of which are calculated based on the SemEval evaluation framework [3, 4].

2. Implementation Details

The implementation builds upon the approach used in the second assignment, with key differences primarily in the data preprocessing and evaluation steps. The dataset is preprocessed by simplifying the labels to just two categories: ['T', 'O'], where T' corresponds to a token that represents an aspect term, and O' corresponds to other tokens. An example of the preprocessing is shown below:

```
Boot=T-POS time=T-POS is=O
super=O fast=O ,=O around=O
anywhere=O from=O 35=O seconds=O
to=O 1=O minute=O .=O
```

This snippet represents an original sentence from the dataset. After preprocessing, the transformed version of the same sentence would appear as:

```
Boot=T time=T is=O super=O
fast=O ,=O around=O anywhere=O
from=O 35=O seconds=O to=O 1=O
minute=O .=O
```

For consistency, the `bert-base-uncased` model and tokenizer are employed, consistent with prior assignments. The tokens are processed using BERT's tokenizer to ensure the input is correctly formatted for the model. Notably, only the first subtoken of a word receives the appropriate tag ID, with all subsequent subtokens being treated as padding (ID = 0). This strategy is essential for avoiding evaluation distortions caused by issues related to sub-tokenization.

2.1. Fine-Tuning

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

3. Results

Table 1 shows the results for the experiments. Training with learning rate scheduling indeed helped with stability, while training with both scheduling and clipping lead to 'noisier' models.

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] X. Li, "E2e-tbsa," https://github.com/lixin4ever/E2E-TBSA/tree/master/data.

[3] ——, "Evaluation script for semeval-2014 task 4," https://github.com/lixin4ever/E2E-TBSA/blob/master/evals.py.

[4] SemEval 2014 Organizers, "Semeval-2014 task 4: Aspect based sentiment analysis - data and tools," 2014. [Online]. Available: https://alt.qcri.org/semeval2014/task4/index.php?id=data-and-tools

Part	**Experiment**	**LR**	**F1** (%)	**Precision** (%)	**Recall** (%)
1.0	Bert	2e-5	97.63 $\pm$0.01	97.64 $\pm$0.01	97.64 $\pm$0.01
1.0	Bert	3e-5	97.57 $\pm$0.02	97.58 $\pm$0.02	97.58 $\pm$0.02
1.0	Bert	5e-5	**97.8** $\pm$0.07	**97.8** $\pm$0.08	**97.8** $\pm$0.05
1.1	BertSch	5e-5	**97.85** $\pm$0.01	**97.86** $\pm$0.01	**97.86** $\pm$0.01
1.2	BertClip	5e-5	97.67 $\pm$0.11	97.67 $\pm$0.08	97.67 $\pm$0.11
1.3	BertSchClip	5e-5	97.85 $\pm$0.18	97.85 $\pm$0.18	97.85 $\pm$0.18

Table 1: *Performance of the models for each configuration.* ***Bold*** *values represent the best model for each section, underline the best overall.*