# A Appendix

## A.1 Technical Details

The model was implemented by using the PyTorch library and based on open-source code (Wolf et al., 2020). For the initial checkpoint for BERT, we adapted the BERT$_{base}$ from Devlin et al. (2019). The optimizer was AdamW optimizer proposed by Loshchilov and Hutter (2019).

In the post-training step, the maximum sequence length was set to 240, and the batch size was set to 50. The learning rate was 3e-5. The number of warm-up steps was 0.01% of the total steps.

When we executed fine-tuning, the maximum sequence length was set to 256, the context and response length ratio was set to 1:1, and the batch size was set to 32. When measuring the performance, we calculated the average values among the five measurements from the different seeds.

## A.2 Hyperparameters

We fine-tuned BERT-FP with an initial learning rate 1e-5, which achieved the highest score, as shown in Table 1. The learning rate was heuristically determined with less than ten trials of adjustment. The model also converged in one epoch. Other hyperparameters are the same as those described in Section A.1. The post-training is stopped when the validation recall does not increase, as shown in Table 2. Specifically, We stopped post-training at 25th, 27th, and 34th epoch for the Ubuntu, Douban, and E-commerce Corpus, respectively. As shown in Table 3-5, We measure the performance of BERT-FP with five different seeds on three benchmarks

## A.3 Datasets link

- Ubuntu Corpus - https://www.dropbox.com/s/2fdn26rj6h9bpvl/ubuntu_data.zip?dl=0
- Douban Corpus - https://www.dropbox.com/s/90t0qtji9ow20ca/DoubanConversaionCorpus.zip?dl=0
- E-commerce Corpus - https://drive.google.com/file/d/154J-neBo20ABtSmJDvm7DK0eTuieAuvw/view

| Learning rate | $R_{10}@1$ | $R_{10}@2$ | $R_{10}@5$ |
|---|---|---|---|
| 5e-6 | 0.860 | 0.933 | 0.986 |
| 1e-5 | 0.911 | 0.962 | 0.994 |
| 2e-5 | 0.900 | 0.957 | 0.993 |

Table 1: Validation performances according to the fine-tuning learning rate on the Ubuntu Corpus.

| # of epoch | $R_{10}@1$ | $R_{10}@2$ | $R_{10}@5$ |
|---|---|---|---|
| 23 | 0.909 | 0.962 | 0.993 |
| 24 | 0.911 | 0.962 | 0.993 |
| 25 | 0.911 | 0.962 | 0.994 |

Table 2: Validation performances according to the number of epochs on the Ubuntu Corpus.

## A.4 Computing infrastructure

Table 7 describes the software used for running the experiment. Table 8 describes the hardware used for running the experiment. Our experiments, both post-training and fine-tuning, are run on a single GPU, Titan RTX 24G.

## A.5 Abbreviation

Table 9 indicate abbreviation.

## References


Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

| Seed | 0 | 2020 | 1118 | 404 | 904 | Average | STD |
|---|---|---|---|---|---|---|---|
| $R_{10}@1$ | 0.91060 | 0.91126 | 0.91188 | 0.91002 | 0.90990 | 0.91073 | 0.00084 |
| $R_{10}@2$ | 0.96236 | 0.96230 | 0.96284 | 0.96268 | 0.96232 | 0.96250 | 0.00024 |
| $R_{10}@5$ | 0.99362 | 0.99420 | 0.99368 | 0.99452 | 0.99386 | 0.99398 | 0.00038 |

Table 3: Validation performances according to seeds on the Ubuntu Corpus.

| Seed | 0 | 2020 | 1118 | 404 | 904 | Average | STD |
|---|---|---|---|---|---|---|---|
| $R_{10}@1$ | 0.91082 | 0.91112 | 0.91082 | 0.91070 | 0.91076 | 0.91084 | 0.00016 |
| $R_{10}@2$ | 0.96222 | 0.96270 | 0.96262 | 0.96130 | 0.96286 | 0.96227 | 0.00070 |
| $R_{10}@5$ | 0.99358 | 0.99410 | 0.99396 | 0.99338 | 0.99380 | 0.99376 | 0.00029 |

Table 4: Test performances according to seeds on the Ubuntu Corpus.

| Seed | 0 | 2020 | 1118 | 404 | 904 | Average | STD |
|---|---|---|---|---|---|---|---|
| $MAP$ | 0.644 | 0.640 | 0.643 | 0.656 | 0.636 | 0.644 | 0.007 |
| $MRR$ | 0.683 | 0.681 | 0.680 | 0.680 | 0.676 | 0.680 | 0.003 |
| $P@1$ | 0.514 | 0.513 | 0.505 | 0.523 | 0.505 | 0.512 | 0.007 |
| $R_{10}@1$ | 0.326 | 0.322 | 0.322 | 0.333 | 0.319 | 0.324 | 0.005 |
| $R_{10}@2$ | 0.543 | 0.528 | 0.549 | 0.570 | 0.520 | 0.542 | 0.019 |
| $R_{10}@5$ | 0.868 | 0.867 | 0.870 | 0.880 | 0.866 | 0.870 | 0.006 |

Table 5: Test performances according to seeds on the Douban Corpus.

| Seed | 0 | 2020 | 1118 | 404 | 904 | Average | STD |
|---|---|---|---|---|---|---|---|
| $R_{10}@1$ | 0.874 | 0.866 | 0.866 | 0.867 | 0.877 | 0.870 | 0.005 |
| $R_{10}@2$ | 0.958 | 0.955 | 0.957 | 0.954 | 0.958 | 0.956 | 0.002 |
| $R_{10}@5$ | 0.993 | 0.992 | 0.994 | 0.994 | 0.993 | 0.993 | 0.001 |

Table 6: Test performances according to seeds on the E-commerce Corpus.

| | |
|---|---|
| Operating system | Linux version 4.15.0-106-generic<br>(buildd@lcy01-amd64-016)<br>(gcc version 7.5.0 (Ubuntu 7.5.0-3ubuntu1 18.04)) |
| Frameworks | Python==3.6.8<br>Anaconda==4.7.12<br>NVIDIA (R) Cuda compiler driver V10.0.130 |
| Libraries | Numpy==1.16.4<br>tokenizers==0.5.2<br>torch==1.5.0<br>torchvision==0.6.0a0+82fd1c8<br>tqdm==4.46.0<br>transformers==2.8.0<br>cudatoolkit==10.1.243 |

Table 7: Software specification.

| | |
|---|---|
| CPU | Intel (R) Core (TM) i7-8700K CPU @ 3.70GHz |
| RAM | 64GB |
| GPU | Titan RTX 24G |

Table 8: Hardware specification.

| | |
|---|---|
| ALBERT | A Lite BERT |
| BERT | Bidirectional encoder representations from transformers |
| BERTbase | Pre-trained BERT base mode(110M , not fine-tuned) |
| BERT-DPT | BERT+Domain post-training |
| BERT-FP | BERT model with fine-grained post-training |
| BERT-FP-NF | BERT+fine-grained post-training+No fine-tuning |
| BERT-SL | BERT with self-supervised tasks |
| BERT-VFT | BERT+Variable fine-tuning |
| DA | Data augmentation |
| DMN | Deep attention matching network |
| DUA | Deep utterance aggregation |
| ESIM | Enhanced sequential inference model |
| FP | Fine-grained post-training |
| IOI | Interaction-over-interaction network |
| MAP | Mean average precision |
| MLM | Masked language mode |
| MRR | Mean reciprocal rank |
| MSN | Multi-hop selector network |
| NSP | Next sentence prediction |
| P@1 | Precision at one |
| $R_{10}$@1 | Recall 10 at 1 |
| RoBERTa | Robustly optimized BERT |
| RoBERTa-SS-DA | RoBERTa+Speaker Segment + Data augmentation |
| SA-BERT | Speaker-aware embedding+BERT |
| SCR | Short context-response pair |
| SMN | Seqential matching network |
| SOP | Sentence ordering prediction |
| UMS | Utterance manipulation strategies |
| URC | Utterance relevance classification |

Table 9: Abbreviation appendix