期末项目 - 中文词性标注

李拙 2001213060
lizhmq@pku.edu.cn

1 任务介绍

词性标注是指在给定句子中判定每个词的语法范畴，确定其词性并加以标注的过程，这也是自然语言处理中一项非常重要的基础性工作。该项目为中文词性标注：给定中文分词结果，需要预测每个分词的词性标签。例如，训练集中的一条数据为：

“海内外/Nl 關注/Vt 的/Us 一九九七/Mo 年/Qc 七月/Nt 一/Mo 日/Qc 終於/Dc 來到/Vt 。/Sy”。

那么，任务的输入为：

“海内外 關注 的 一九九七 年 七月 一 日 終於 來到 。”，

需要预测的输出则为：

“Nl Vt Us Mo Qc Nt Mo Qc Dc Vt Sy”。

该项目分两部分，分别为简体中文的词性标注和繁体中文的词性标注。我们对两个任务各训练了一个配置相同的模型。

2 实验设置

2.1 模型设置

我们使用 Chinese-RoBERTa-wwm-ext(1) 作为基础模型，用其对句子中的每一个字产生上下文相关的嵌入向量表示，然后再用线性层对该嵌入向量进行分类[1]。

Chinese-RoBERTa-wwm-ext BERT(2) 是 Google 在 2018 年提出的预训练模型，其核心组件是基于注意力机制的 Transformer(4) 模型。BERT 在大规模的语料上通过使用 Masked Language Model Prediction 和 Next Sentence Prediction（NSP）来进行预训练。通过 BERT 预训练学习到的文本特征能够有效地迁移到下游任务上，在一系列的自然语言处理任务上取得了 SOTA 表现。RoBERTa(3) 是 BERT 的改进版，其作者通过大量实验，得到了训练 BERT 类模型的通用经验性方法：比如去除 NSP 任

[1]注：Chinese-RoBERTa 对每个字产生向量表示，我们预测一个词的词性时，直接使用第一个字的向量表示。

务，使用更大的 batch size 等。RoBERTa 在很多任务上取得了比 BERT 更好的结果。BERT 和 RoBERTa 的初始工作都在英文预料上训练，Chinese-RoBERTa-wwm-ext(1) 是对 BERT 类模型在中文预料上特别优化的模型。作者考虑了中文特有的分词性质，在训练 Chinese-RoBERTa 时使用了整词掩码（Whole Word Masking - wwm）来提升模型在中文语料上的表现。

2.2 模型参数

Chinese-RoBERTa-wwm-ext 的模型大小与 RoBERTa-base 相同。它由 12 层隐层维度为 768 的 Transformer 层构成，额外添加的线性分类层参数大小为 $768 \times |\mathcal{Y}|$，其中 $\mathcal{Y}$ 为标签空间。整个模型的参数大小约为 97.6M。

2.3 训练细节

我们将两个数据集都按照 8:1:1 划分为训练集、验证集和测试集，在两个训练集上各训练一个模型，最大轮数为 20。训练时 batch size 设置为 32。我们的代码开源在 Github：`https://github.com/Lizhmq/POSTagging-BERT`。

3 实验结果

我们的实验结果列在表1和表2中。详细的各类准确率在 GitHub 中列出。

表 1: 实验结果：数据集一（繁体）

	Accuracy	Macro F1	Average F1
Model	0.95	0.85	0.95

表 2: 实验结果：数据集二（简体）

	Accuracy	Macro F1	Average F1
Model	0.75	0.89	0.97

References

[1] Y. Cui, W. Che, T. Liu, B. Qin, Z. Yang, S. Wang, and G. Hu. Pre-training with whole word masking for chinese BERT. *CoRR*, abs/1906.08101, 2019. URL `http://arxiv.org/abs/1906.08101`.

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.