计算语言学 Project
深度学习在中文分词方面的应用

lzh

日期：2020 年 12 月 25 日

1 研究动机

中文分词在中文自然语言处理里算是一个比较重要的任务，虽然近年来的自然语言处理相关的神经网络工作基本都是基于中文字符，但是在搜索领域还是大量采用中文分词加上匹配的方法对 query 进行处理，所以中文分词依旧是一个比较重要的工作。此外，深度学习的发展也给传统的自然语言处理带来新的思路，本次我主要探究深度学习方法在中文分词领域上的工作，并利用 bert 预训练模型实现一个简单的 demo。

2 中文分词问题定义

在深度学习领域，中文分词问题可以定义为：给定中文句子序列 $c_1c_2...c_n$，对其进行序列标注，得到字符序列的分类结果序列 $y_1y_2...y_n$，其中 $y_i \in [B, E, M, S]$，其中 B 代表该字符为词的开始字符，E 代表该字符为词的结束字符，M 代表该字符在一个词的中间，S 代表该字符单字成词。

3 分词算法

有了如上思路，可以很简单地采用序列标注的模型对该问题进行建模。在本次探究中，由于分词语料的数量并不算多，所以考虑采用最近特别火的预训练模型 BERT 作为基础模型，在预训练模型上利用分词语料进行微调。假设例句为“我不学习了！”，输入 BERT 前，该句子前后需要加入特定标识符 [CLS] 与 [SEP]，随后 BERT 对其进行处理得到该序列的表示向量 $[t_{CLS}, t_1, t_2, ..., t_6, t_{SEP}]$，利用分类网络对得到的每个字符的表示向量进行分类，即可完成中文分词任务。模型结构图如下：

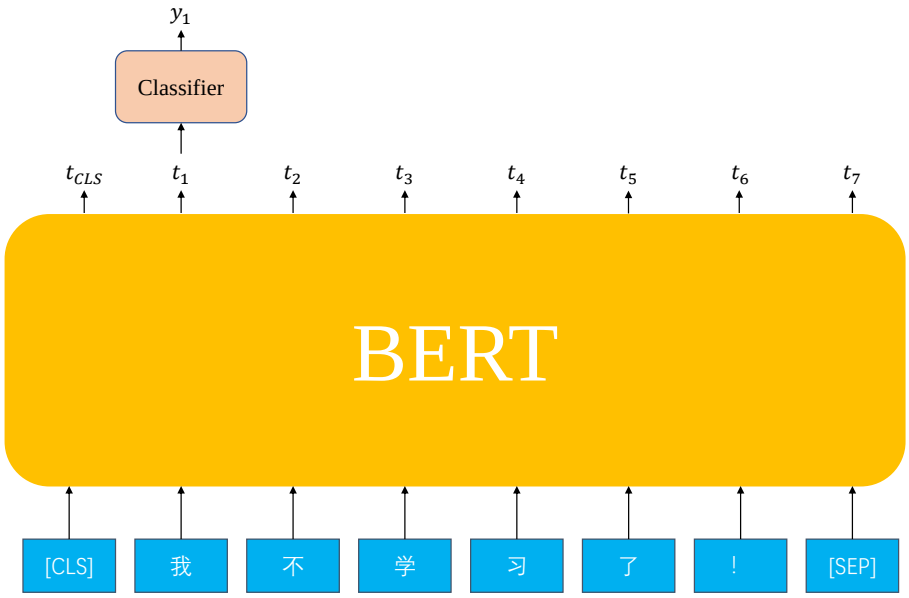

图 1: BERT 中文分词模型

4 数据集以及预处理

数据集上，选用The Second International Chinese Word Segmentation Bakeoff中提供的四个中文分词语料，分别为：

- Academia Sinica
- CityU
- Peking University
- Microsoft Research

为了使得训练语料尽可能多，将四个分词语料合并进行训练。预处理方面进行以下操作：

1. 将 CityU 与 Academia Sinica 语料中的繁体字转换为简体字
2. 将所有语料中的全角字符转换为半角字符
3. 所有英文字母变为小写
4. 对于语料中的每个词，对其进行中文字符级别的切分，保留英文单词以及数字
5. 切分完毕后，生成序列对应的分词标注，并保存到一个 json 文件中

除此之外，在实际上手时，发现 bert 的 tokenizer 会对英文单词进行进一步的拆解，为了简便起见，对于该拆分的单词，其对应的标注方法同中文词。由于语料中英文单词的比重较少，该操作并不会对准确率造成较大影响。

5 实验结果

实验中采用 huggingface 提供的 bert-base-chinese 模型作为预训练模型，训练过程中采用 batchsize=16，优化器为 Adam，bert 网络学习率设为 1e-5，分类器部分学习率为 1e-2。训练集与

验证集比例为 9:1，训练 8 个 epoch 并保存验证集 loss 最小时的模型。除此之外，由于显存有限，将最大的句子长度设置为 300（数据集中长度大于 300 的句子只有约 800 条，远小于总数据量，可以忽略）。分类结果如下：

表 1: 两层分类器下分类结果

	precision	recall	f1-score	support
BEGIN	0.98	0.97	0.97	189989
END	0.97	0.97	0.97	190108
MIDDLE	0.87	0.91	0.89	39221
SINGLE	0.96	0.96	0.96	171511

表 2: 单层分类器下分类结果

	precision	recall	f1-score	support
BEGIN	0.97	0.98	0.97	189989
END	0.97	0.97	0.97	190108
MIDDLE	0.89	0.88	0.88	39221
SINGLE	0.97	0.95	0.96	171511

可以看出 Bert 的分词器效果还是很好的，但是在字符为 MIDDLE 时的判定表现较差，可能的原因是分词数据中大部分词都是单字或者两个字组成的，长于 2 的词出现频度没有那么高。在运行效率方面，实际速度并不算慢，单论模型计算，在 GPU 上测试整个测试集，耗时为 2 分 33 秒左右。

下面列举几个分词实例，对比了 bert 分词、jieba 分词与 thulac 分词效果：

```
例句：中国青年报发表文章《做题家们的怒气为何要往丁真身上撒》
Bert：中国青年报 发表 文章 《 做题家 们 的 怒气 为何 要 往 丁真 身上 撒 》
jieba：中国青年报 发表文章 《 做题 家们 的 怒气 为何 要往 丁真 身上 撒 》
thulac：中国 青年报 发表 文章 《 做 题 家 们 的 怒气 为何 要 往 丁真 身上 撒 》
```

可以看出上述句子中，“做题家”是一个新词，bert 可以识别出来，jieba 就分词错误，thulac 则将“做题家”切的太细，所以在这里也可以看出，中文分词实际上是存在多种分法的，并不好评价分词的正确与否。

```
例句：内卷、打工人、做题家、小丑竟是我自己都是近年来较火的自嘲梗
Bert：内卷 / 、 / 打工人 / 、 / 做 / 题家 / 、 / 小丑 / 竟是 / 我 / 自己 / 都 /
    是 / 近年 / 来 / 较 / 火 / 的 / 自嘲梗
jieba：内 / 卷 / 、 / 打 / 工人 / 、 / 做题 / 家 / 、 / 小丑 / 竟是 / 我 / 自己
    / 都 / 是 / 近年来 / 较火 / 的 / 自嘲 / 梗
thulac：内卷 / 、 / 打工 / 人 / 、 / 做 / 题 / 家 / 、 / 小丑 / 竟是 / 我 / 自
    己 / 都 / 是 / 近年 / 来 / 较 / 火 / 的 / 自嘲梗
```

上述例子中，Bert 就错误地将“做题家”切为“做题家”，jieba 则将“打工人”切分为“打工人”，这就政治不正确了。thulac 总体来说粒度较细，准确率不错。

6 总结

本次实验测试了 Bert 在中文分词领域的运用，可以很明显地体会到预训练模型在 NLP 领域的威力。此外，基于神经网络的中文分词，基本都是采用序列标注的方法进行计算，但是序列标注只是一个比较简单直观的指标，而且分词的正确结果并不统一，所以局限性还是比较大的，但是要做出一个可以用的中文分词模型却是简单的。