

**Harbin Institute of Technology**

**商品标题实体识别**

(2022春 自然语言处理 课程报告)

姓名: 肖力炜 学号: 190110504

姓名: 王铭 学号: 190110509

姓名: 郑羿恺 学号: 190110502

4/22/2022

**Contents**

[1. 引言 1](#_Toc101566848)

[1.1 1](#_Toc101566849)

[2. 数据分析 1](#_Toc101566850)

[2.1 1](#_Toc101566851)

[3. 模型设计及实现 1](#_Toc101566852)

[3.1 1](#_Toc101566853)

[3.2 1](#_Toc101566854)

[4. 实验结果 2](#_Toc101566856)

[4.1 A榜排名及分数 2](#_Toc101566857)

[4.2 B榜排名及分数 2](#_Toc101566858)

[5. 组员及分工 2](#_Toc101566859)

[参考文献 3](#_Toc101566860)

1. 引言

1.1 NER的定义与应用

命名实体识别（Named Entity Recognition，简称NER）是自然语言处理中的热点研究方向之一，目的是识别文本中的命名实体，并将其归纳到相应的实体类型中。命名实体识别也是NLP最重要的底层任务之一，在学术界和工业界一直都是重点研究的问题。

NER就是识别文本中具有特定意义的实体，在商品标题识别中，主要包括商品名称、厂家、型号、参数、颜色、适宜人群等，如：

Apple iPhone 13 (A2634)128GB 星光色 支持移动联通电信5G 双卡双待手机

其中，“iPhone”为商品名称，“13”为型号，“Apple”为厂家，“128GB”为参数，“星光色”为颜色，“支持移动联通电信5G 双卡双待”为适宜人群，“手机”为商品类别。

再如：

阿诺顿【3件装】短袖t恤男夏季新品男士T恤大码宽松中青年潮流男生打底衫半袖上衣服 PK黑+KY7深灰+音乐雾蓝 L 120-135斤

包括商品名“短袖t恤”“T恤”“打底衫”“半袖”“上衣（服）”，季节“夏季”，适宜人群“男士”“中青年”“男生”，型号“大码”“宽松”“L”“120-135斤”，颜色“PK黑”“KY7深灰”“音乐雾蓝”等信息。

NER在多种类型的文本上都有着广泛的应用，主要分为两大类：用户query理解和商品标题文本结构化。用户的query理解，包括了搜索框内的文本搜索以及语音场景下的搜索。比如说这里用户在搜索框内输入的query是“刻晴cos服霓裾翩跹”，我们库内并没有同名资源，需要使用NER模块从中提取出来商品类别“cos服”，相关人物“刻晴”以及商品名“霓裾翩跹”，这时候系统才知道用户想要一件原神主题人物刻晴的一套名为“霓裾翩跹”时装同款衣服。商品标题文本结构化则主要是通过商家售卖产品的标题中抽取一些实体，便于在搜索推荐等任务中的应用。

1.2：NER发展历史

早期的NER基于规则和词典，还有统计机器学习的方法（HMM、CRF、SVM等），主要的优点是速度快，精确率比较高，但是召回率通常比较低。目前，面对搜索场景下大多数请求中的高频query的快速相应机制会使用这类方法。

2016~2018年之间，随着深度学习的发展，基于浅层神经网络的方法开始应用于NER领域。它的基础结构是通过LSTM或者IDCN去抽取文本的特征，再通过CRF解码，得到对应的实体序列，在此基础上，融入外部词典特征和相关领域内信息等一些可以被参考的外部知识进行模型优化。此外，一些多粒度信息可以被选中加入模型，比如LSTM+CHAR-CNN增加一个字符级的CNN融入，LATTICE-LSTM通过对LSTM的结构进行修改实现词以及短语等不同粒度信息的融入。这类方法目前依然是工业界主流的解决方案。

2018年后，随着Bert的兴起，基于大规模预训练语言模型的方法成为了趋势，这类方法主要特点是，使用Bert这样的大规模训练语言模型代替浅层的文本抽取器、特征提取器，来获取高质量的embedding去优化下游的任务。此外在这类方法的基础上还会有一些相关的优化，用于实现外部词典信息的引入。

1.3：NER的难点

（1）：领域相关性强。如果没有领域内的相关知识，文本中的实体难以正确地被识别，比如：“明日方舟纪念插画集官方艺术美术设定集Vol.2第2弹”，“明日方舟”作为上海鹰角网络发行的手游，应当被识别成一个单词，然而“明日”“方舟”都可以单独成词，若没有对应的知识，仅文本层面明日方舟可能会被理解成明天的方舟之类的普通单词，就失去了描述该商品最重要的属性。

（2）：实体命名歧义大。一个实体名在不同场景下实体类型可能不同。比如用户单独搜索“苹果”，我们无法确定用户想买苹果这种水果还是Apple的电子产品。比如搜索“中华”，可能指中华香烟、中华牙膏、中华书局等等。

（3）：缺乏足够的上下文信息。商品标题通常较短，或者是各类关键词堆叠以期用户能够检索到，在划分实体时若无上下文参考，容易误判实体名称边界。这种情况在“三坑”类别商品 中十分常见。如“一只冲天兔绒绒云”，“一只冲天兔”为店铺名，绒绒云为商品名。然而，分词时容易分出“兔绒”这一词，发生边界错误。另一例子则是“雾岛云辉月海南国暮蝶”，这类商品名毫无上下文关联可以利用，给NER工作造成了较大困难。

（4）：表述具有多样性。开放域文本具有多样性，同样的意思可能有多种表达方式，并且表达方式相对于正规文本来说更加的不规范。

2. 数据分析

2.1数据预处理

该任务提供的原数据如图所示：

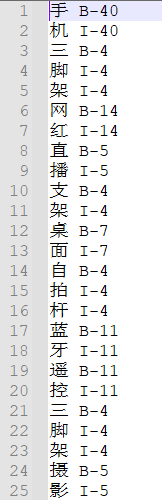


图2.1 原训练集数据

为了后面的训练、预测，可以对数据做预处理，将单个字及其对应的标签转换为对应的句子及标签列表，观察原数据知，换行符代表一句话的结束，则两个换行符之间的应组成一个句子，同时考虑空格的特殊情况：

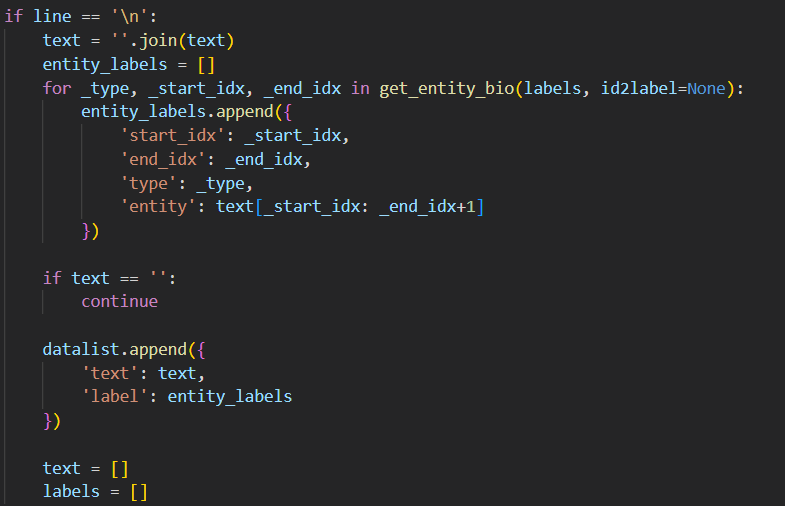


图2.2-1 数据处理部分

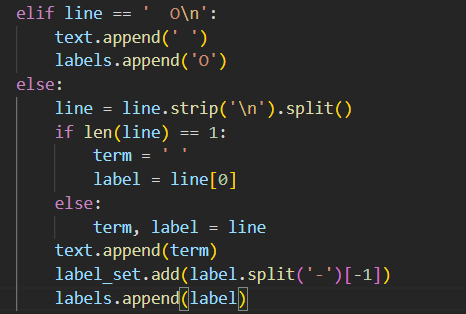


图2.2-2 数据处理部分

处理后的结果如图所示：

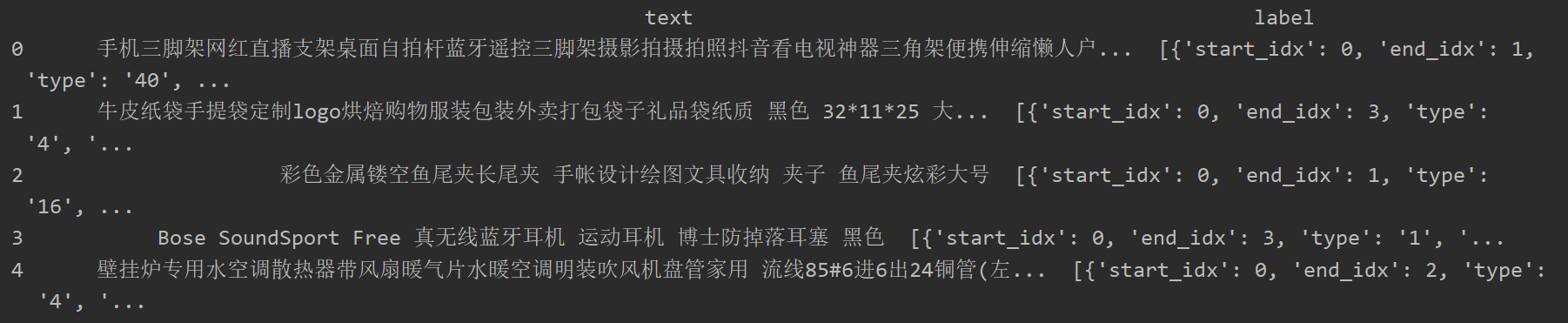


图2.3 数据处理结果

2.2训练集与验证集划分

简单的划分数据集和验证集，用整个训练集的后400条数据作为验证集（这里选择直接用训练集的后400条数据作为验证集是因为采用sklearn的交叉验证集划分后，效果反而在下降）

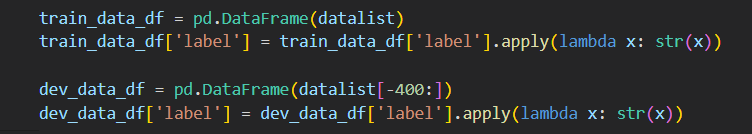


图2.4 训练集与验证集划分

采用globalpointerbert的数据集加载划分后的训练集与验证集

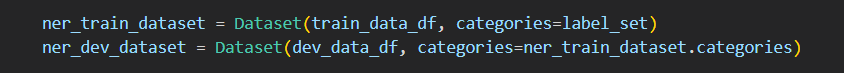


图2.4 加载数据集Dataset

2.3创建词典、分词与ID化

NER任务中，输入的原数据为自然语言，训练模型时，应该将自然语言编码（或映射）成数字，即ID化再进行训练，这里词典直接使用开源的chinese-bert-wwm，wwm即Whole Word Masking，该模型会对被选中mask的词进行全词mask，该tokenizer直接实现了对文本的分词、ID化、填充对齐等操作。

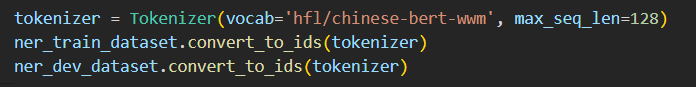


图2.5 创建词典并进行分词、ID化

3. 模型设计及实现

3.1 模型创建

创建的模型采用Bert+GlobalPointer模型，该模型利用全局归一化的思路来进行命名实体识别，与传统模型在训练时分别从实体的首、尾识别导致在预测未知数据时的不一致相比，使用该模型可以更好地掌握全局信息，同时消除了训练和预测时出现的不一致问题。

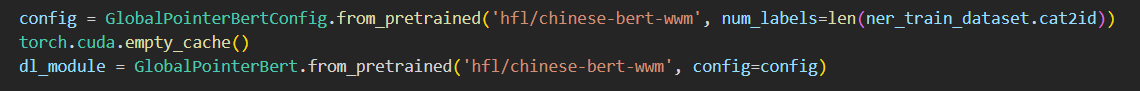


图3.1基于Bert+GlobalPointer创建模型

3.2 模型训练

优化器选择默认优化器，创建任务使用globalpointerbert的Task，学习率lr设置为常用的2e-5，在尝试了几个训练轮次后在最终的测试集上预测，选择训练轮数为5轮，从而避免过拟合。

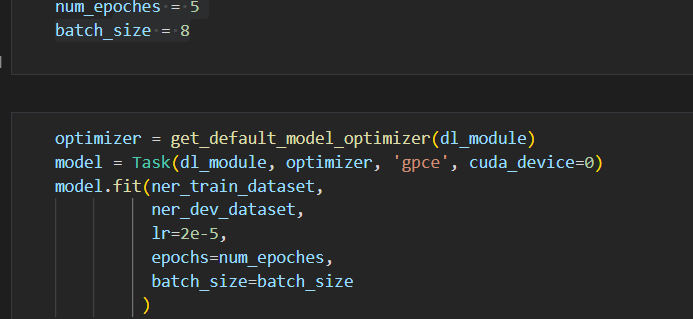


图3.2模型训练

3.3 模型预测

预测采用GlobalPointer命名实体识别的预测器Predictor进行预测，由于训练时将自然语言转换成了ID（即encode过程）,该模型同样先进行ID化再利用训练好的模型预测实体类别，同时完成解码。

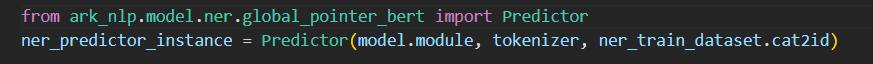


图3.3构建预测模型

通过predict\_one\_sample方法进行预测，对预测结果进行处理，主要处理为实体开始位置的标签应该为B，中间全为I，对开头不是B等发生了嵌套情况的标签全设置为O。

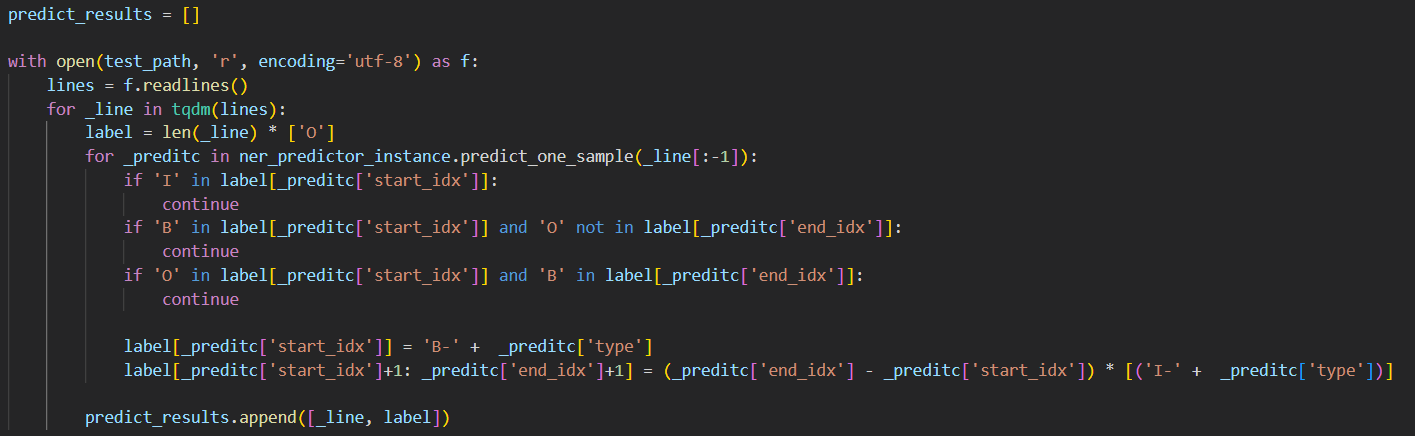


图3.4 NER实体预测

4. 实验结果

4.1 A榜排名及分数



4.2 B榜排名及分数



5. 组员及分工

|  |  |  |
| --- | --- | --- |
| 姓名 | 学号 | Task Assignment (code and report) |
| 王铭 | 190110509 | Programming, Data processing, Report |
| 肖力炜 | 190110504 | Algorithm, Data analysis, Report |
| 郑羿恺 | 190110502 | Programming, Report |

参考文献

1. Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//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). 2019: 4171-4186.