基于Bert的关系抽取和实体识别

目录：

一、总体简介

1.1 任务介绍

2019 CCF大数据与计算智能大赛（训练赛-文本实体识别及关系抽取），利用经过特定处理的公共数据集SemEval2010，数据集中的文本共包括9种实体关系，希望参赛者

- 对句子进行实体抽取
- 并根据语义及其他信息来判断实体之间的关系

1.2 数据集介绍

数据集中一共有10717条英文文本，分为8000条训练数据以及2717条测试数据，包含9种实体间的关系，数据集文件也有对9种实体关系的详细描述及距离，帮助参赛者更深入的理解各种实体关系的含义。

数据格式：

“The system as described above has its greatest application in an arrayed configuration of antenna elements.”

Component-Whole(elements, configuration)

其中第一句是文本信息，参赛者应当分析这个句子，提取出句子中的实体，这些句子都包含一对以上的实体，因此无需考虑句子中只包含单实体的情况。

第二句Component-Whole(elements, configuration)是得到的结果，提取到的实体对是(elements, configuration)，实体之间的关系是Component-Whole。当句子中包含多对实体时，我们仅要求参赛者的判断结果中包含我们的提供的结果即可。

实体关系类型：

Other、Cause-Effect、Component-Whole、Entity-Destination、Product-Producer、Entity-Origin、Member-Collection、Message-Topic、Content-Container、Instrument-Agency

1.3 实验环境

- TensorFlow 1.12
- python 3.6 +
- Bert-base

二、数据预处理

2.1 用正则表达式处理

原句：

```
The factory 's products have included flower pots, Finnish rooster-whistles, pans, trays , tea pots, ash trays and air moisturisers.
```

新句：

```
the factory products have included flower pots finnish rooster - whistles pans trays tea pots ash trays and air moisturisers
```

2.2 去重

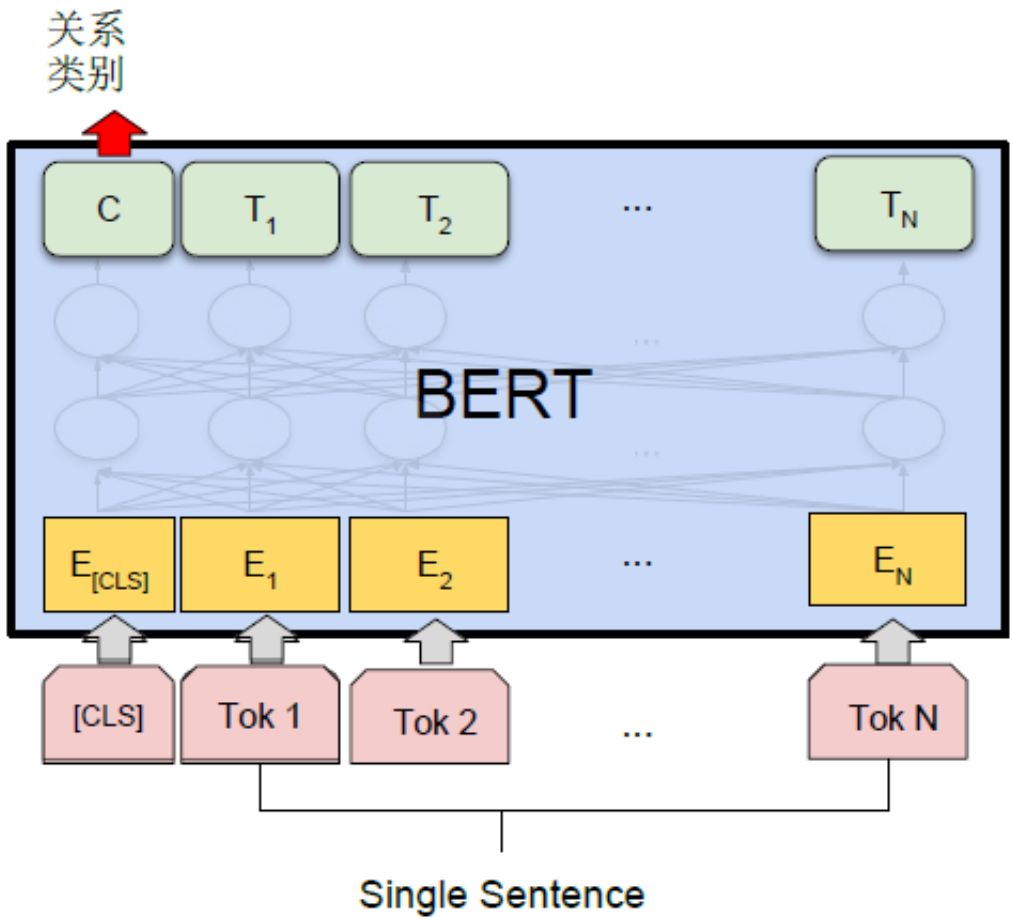
数据分布：训练集8000句中，有25对是重复的句子，这些句子含有多对实体对，多对关系。

策略：由于重复句子数据稀疏，删除重复句子对实验结果影响不大。

三、级联模型介绍

3.1 关系抽取

关系抽取看做是一个分类任务




```
guid: train-0
tokens: [CLS] the system as described above has its greatest application in an
input_ids: 101 1996 2291 2004 2649 2682 2038 2049 4602 4646 1999 2019 9140 2098
input_mask: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0
segment_ids: 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
label: Component-Whole (id = 6)
```

3.2 实体识别

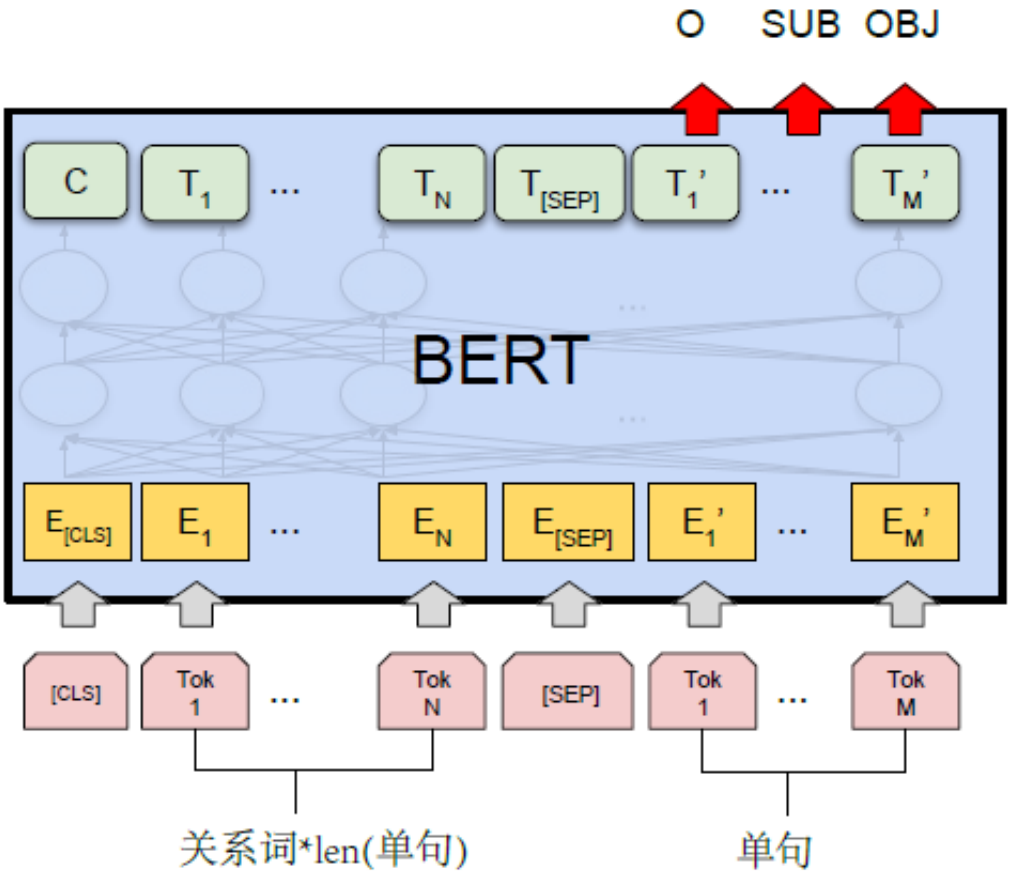
实体识别看做是一个序列标注任务,**关系词作为特征**

输入：与句子等长个数的关系词 +句子

- 训练阶段：关系词为**正确的标注训练数据**
- 测试阶段：关系词为**关系分类模型的预测结果**

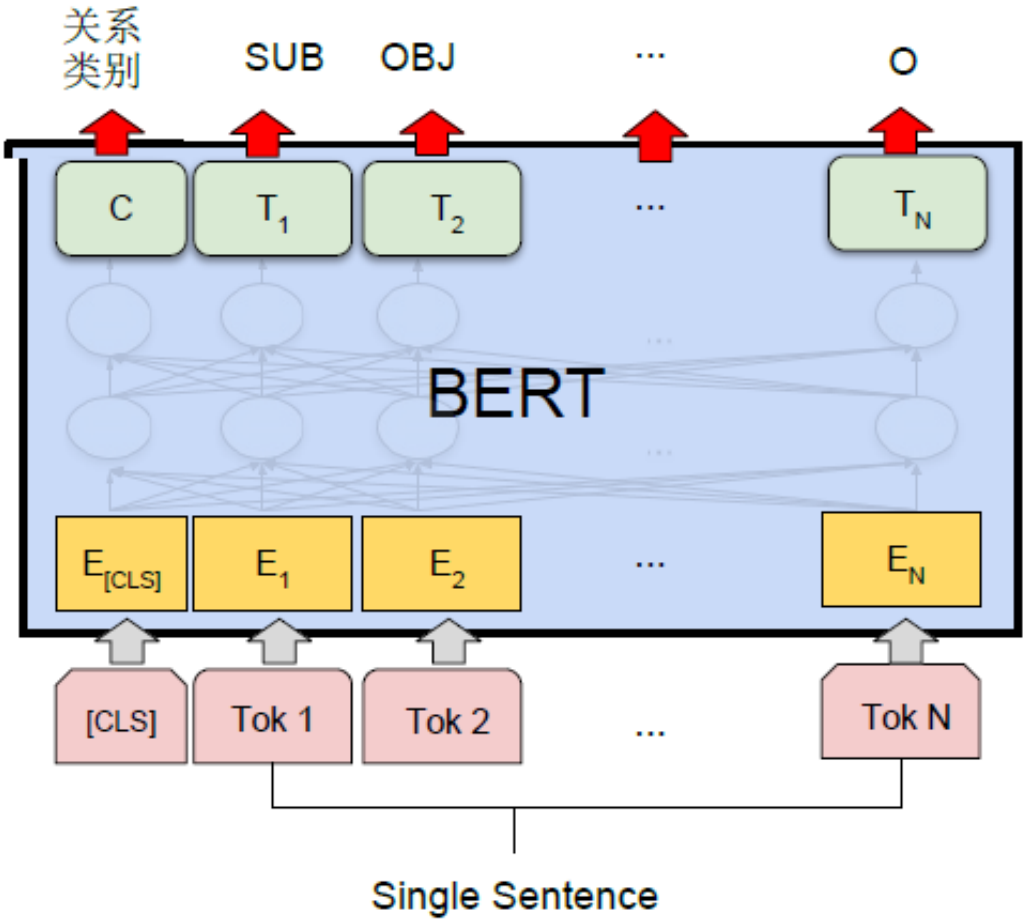
输出：句子对应的序列lable（SUB：主体词，OBJ客体词）

损失函数：交叉熵 or CRF损失函数（2者结果相当）




```
guid: train-0
tokens: the system as described above has its greatest application in an array
input_ids: 101 1996 2291 2004 2649 2682 2038 2049 4602 4646 1999 2019 9140 2098
input_mask: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
segment_ids: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
label_ids: 7 3 3 3 3 3 3 3 3 3 3 3 6 4 3 3 1 8 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9
```

四、联合训练模型介绍




```
 guid: train-0
tokens: the system as described above has its greatest application in an array
input_ids: 101 1996 2291 2004 2649 2682 2038 2049 4602 4646 1999 2019 9140 2098
input_mask: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
segment_ids: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
label_ids: 5 2 2 2 2 2 2 2 2 2 2 2 4 3 2 2 1 6 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
INFO:tensorflow:predicate_label: Component-Whole (id = 6)
```

五、数据后处理

5.1级联模型

抽取对应的实体。如下图为（**audit,waste)**

```
the O O
most O O
common O O
audit B-SUB B-SUB
were O O
about O O
waste B-OBJ B-OBJ
and O O
recycling O O
[SEP] [SEP] [SEP]
Message-Topic [category] [PAD]
Message-Topic [category] [PAD]
Message-Topic [category] [PAD]
Message-Topic [category] [PAD]
Message-Topic [category] [PAD]
Message-Topic [category] [PAD]
Message-Topic [category] [PAD]
Message-Topic [category] [PAD]
Message-Topic [category] [PAD]
Message-Topic [category] [PAD]
[SEP] [SEP] [SEP]
```

5.2 联合训练模型

```
the O O
most O O
common O O
audit B-SUB B-SUB
were O O
about O O
waste B-OBJ B-OBJ
and O O
recycling O O
[SEP] [SEP] [SEP]
```

六、实验结果

6.1 级联模型

（1）关系分类模型

maxLen=32

iteration	eval_accuracy
100	0.74
200	0.75
300	0.78
400	0.79
500	0.79
600	0.79

（2）实体识别模型

SUB,OBJ为正例

maxLen=64

iteration	f1	precision	recall
100	0.74	0.83	0.77
200	0.84	0.84	0.85
300	0.86	0.85	0.87
400	0.87	0.85	0.88
500	0.87	0.86	0.88
600	0.88	0.87	0.88

6.2 联合训练模型

iteration	类别accuracy	f1	precision	recall
200	0.76	0.76	0.80	0.73
400	0.78	0.81	0.82	0.80
498	0.79	0.81	0.82	0.81

6.3、比赛结果

级联模型：

accuracy=0.33,排名10/1572(12月30日排名)

联合模型：

accuracy=0.28,排名11/1572(12月30日排名)