提取聊天对方的称谓

1. 任务目标

通过对双方对话内容的分析，提取出聊天对方的“称谓”，比如说“龙哥”，“韬哥”，“老板”等，然后将聊天人的称谓与姓名相匹配（e.g. 张玉龙 is 龙哥）。

e.g. 输入 1. 贾舒越

2. 小贾，烦劳帮我取下快递。

输出 贾舒越 is 小贾

注意 输入 1. 贾舒越

2. 璐姐，烦劳帮我取下快递。（张舒越，烦劳帮我取下快递。）

输出 There is no match for 贾舒越

A．问题的分析：本任务属于序列标注任务中的**命名实体识别（Named Entity Recognition, NER）**，学术上 NER 所涉及的命名实体一般包括 3 大类（实体类，时间类，数字类）和 7 小类（人名、地名、组织机构名、时间、日期、货币、百分比）。此任务中需要做人名/称谓的命名实体识别，然后通过规则来做姓名和称谓的匹配。

B．问题的难点：

i. 如何更准确更全面地将人名与称谓从句子中识别出来

ii. 通过实现全称人名与称谓语义对应，而不是对应到了第三方（张冠李戴）。

C. 应用场景与领域适应性：聊天/对话领域

D. 语料标注数据选择：全称大名语料+聊天对话时的称谓语料。

2. 预先的条件：

（1） 默认输入 2（句子）中可能存在着人名称谓，也可能没有。

（2） 默认输入 2（句子）在一定的长度范围之内。

（3） 大名：这里指聊天人的大名，例如：贾舒越、张玉龙、林诗璐等。

（4） 称谓：这里指聊天人的称谓，例如：小贾、龙哥、璐姐等。

（5） 暂时不在数据集里加入特殊别名、别称等称谓比如说“伏地魔”、“胆小鬼”等。

3. 任务方案：

第一步：首先通过 BERT + BiLSTM-CRF 实现人名（这里主要指人的大名与聊天称谓）NER 命名实体识别。

作用：训练后的模型可以将句子中的人名识别出来并输出。

注意：通过此方法不仅仅可以做人名实体识别，即将一句话中的所有出现的人名识别提取出来，还可以做其他名称的识别，比如说地名、组织名等等。（有什么数据集就可以做什么识别。）

1. 在识别的过程中发现，全称人名（e.g. 贾舒越、张玉龙、林诗璐等）可以识别出来，半全称人名（e.g. 舒越、文韬、振聪、诗璐等）也可以识别出来**（只要是人名中的一部分均可以识别出来）**，此类基于 BERT 的 NER 较为准确可靠。
2. 遇到特殊的“小贾”、“璐姐”、“龙哥”、“张大哥”、“李老师”、“戴老板”由于在数据集中加入了相关的聊天称谓标注数据，可以较好地有效识别提取出来。
3. 本方法基于 BERT-base：chinese_L-12_H-768_A-12 预训练模型做 Fine tuning：

```
{
  "attention_probs_dropout_prob": 0.1,
  "directionality": "bidi",
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "max_position_embeddings": 512,
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "pooler_fc_size": 768,
  "pooler_num_attention_heads": 12,
  "pooler_num_fc_layers": 3,
  "pooler_size_per_head": 128,
  "pooler_type": "first_token_transform",
  "type_vocab_size": 2,
  "vocab_size": 21128
}
```

具体配置如上。

4. 本方法使用了数据集有 30676 个标注人名。
5. 本方法 Fine tuning 默认的话迭代次数为 69646 次，所需时间为 10 个小时左右（GTX1060-6G 显卡）。
6. 本方法基于 TensorFlow 1.13.1 + CUDA10.0。
7. 本方法 CSDN 教程为 https://blog.csdn.net/macanv/article/details/85684284
8. 本方法 GitHub 原代码与教程为 https://github.com/macanv/BERT-BiLSTM-CRF-NER

第二步：通过编写规则实现匹配。

编写简单脚本加载模型对输入的句子进行人名 NER 命名实体识别后，得出识别出来的人名与称谓，接下来通过规则实现输入的全称大名与识别人名进行匹配（e.g.小贾 is 贾舒越、There is no match for 张玉龙）。

规则制定：

1） 此规则**基于"字"层面**用于解决话中的称谓与全称人名的匹配对应。

P.S. 此规则很大程度在于 NER 识别的准确性，详情请看错误原因与分析 1：NER 识别人名或称谓 识别错误或者识别不出来

任意称谓中出现了姓名中的字，例如：小龙中的"龙"出现在"张玉龙"中。认为"小龙"是"张玉龙"的一个称谓。即只要有输入的人名和识别出来的人名有一个或多个字相同，就认为识别出来的人名是输入的一个人名的代称。

方法（举例）：输入一："小龙你最近过得怎么样呢？"

输入二："张玉龙"

1. 对输入一，通过第一步的 NER 识别，得出识别出的人名：["小龙"]。
2. 分别对于输入一和输入二进行"字"拆分，得出：

输入一的人名：["小"，"**龙**"] （1）

输入二的人名：["张"，"玉"，"**龙**"] （2）

3. 对比（1）与（2）由于二者均有"**龙**"，则认定"小龙"为"张玉龙"的一个称谓。

2） 此规则将部分识别称谓**合并**，间接提高模型识别准确率。

P.S. 此规则也有些缺陷，详情请看错误原因与分析 3：规则匹配时候导致错误。

识别出"*老师"、"*局长"、"*校长"、"*大哥"最好，但是如果识别的是["郝"，"老师"]这种将一个整体称呼识别拆分为了 2 个，就将 2 者合并为一个，即["郝"，"老师"]变为["郝老师"]。（具体有："老*"、"*老师"、"*老板"、"*处长"、"*局长"、"*将军"、"*司令"、"*主席"、"*先生"、"*女士"、"*师兄"、"*师姐"、"*同学"、"*校长"、"*兄"、"*弟"、"*姐"、"*妹"）

3） 此规则用于解决基于"字"层面匹配时出现的各种**特殊情况**。

P.S. 此时处理的是错误原因与分析 2：后期人名匹配张冠李戴

匹配中可能出现 "最近*戴宇航*学习怎么样啊？"+"*戴海生*"，（两者均有*戴*字，但二者不是同一人）或者 "*张舒越*最近在忙什么呢？"+"*贾舒越*"。（两者均有舒越，但二者不是同一人）

此时出现问题的原因主要是基于"字"层面的匹配←单独的"字"或"词"并无法体现名称间的语义信息。

4. 测试集分布、数量与准确率、错误结果分析

训练后的模型总准确率 Accuracy：0.9673

（1） 测试集共有 550 条，识别正确 531 条，错误 19 条

测试集总共分为四大部分：

1. 含有称谓的 data + 正确的全称大名 ：共 200 条，**错误 15 个**
2. 含有称谓的 data + 错误的全称大名 ：共 200 条，**错误 4 个**
3. 不含称谓的 data + 全称大名 ：共 100 条，**无错误**
4. 不含称谓但有全称大名相同字的 data +全称大名 ：共 50 条，**无错误**
5. 含与全称大名有相同字的人名 data+全称大名 ：共 10 条，**错误 1 个**

（2） 错误的原因与分析：

1. NER 识别人名或称谓 识别错误或者识别不出来-错误 11 条

e.g. 1. “哈哈哈乔大哥”中没有识别出“乔大哥”为称谓。

2. “耿博士，准备去哪所大学呢？”中没有识别出“博士”来。

解决方案：通过对原始数据集再加入几十条与之先关的名称，比如说“博士”、“*总”等，可以识别出这些称呼。**

```
The Print Sentence is  哈哈哈乔大哥
Stage One: Find the -> Title <- in the paragraph: []
The Print Person Name is 乔琪
Print the Model Final Result: There is no match for 乔琪
Current score is  11
Test  12
```

```
The Print Sentence is  最近工作进展怎么样了世昌
Stage One: Find the -> Title <- in the paragraph: []
The Print Person Name is 李世昌
Print the Model Final Result: There is no match for 李世昌
Current score is  50
Test  52
```

```
The Print Sentence is  日呼，是不是在内蒙没有雾霾呢？
Stage One: Find the -> Title <- in the paragraph: []
The Print Person Name is 格格日呼
Print the Model Final Result: There is no match for 格格日呼
Current score is  74
Test  78
```

```
The Print Sentence is  一航，咋们今天晚上一起出来吃顿饭吧。
Stage One: Find the -> Title <- in the paragraph: []
The Print Person Name is 杨一航
Print the Model Final Result: There is no match for 杨一航
Current score is  75
Test  80
```

```
The Print Sentence is  世钍，考研准备怎么样了？
Stage One: Find the -> Title <- in the paragraph: []
The Print Person Name is 张世钍
Print the Model Final Result: There is no match for 张世钍
Current score is  90
Test  97
```

```
The Print Sentence is  耿博士，准备去哪所大学呢？
Stage One: Find the -> Title <- in the paragraph: ['耿']
The Print Person Name is 耿靖童
Print the Model Final Result: 耿 is 耿靖童
Current score is  302
Test  314
```

```
The Print Sentence is  马总，你觉得字节跳动这家公司怎么样呢？
Stage One: Find the -> Title <- in the paragraph: ['马']
The Print Person Name is 马立明
Print the Model Final Result: 马 is 马立明
Current score is  307
Test  321
```

```
The Print Sentence is  祝你生日快乐！世昌
Stage One: Find the -> Title <- in the paragraph: []
The Print Person Name is 李世昌
Print the Model Final Result: There is no match for 李世昌
Current score is  330
Test  346
```

```
The Print Sentence is  王浩大哥，在ABB公司你的岗位是什么呢？
Stage One: Find the -> Title <- in the paragraph: ['王浩']
The Print Person Name is 王浩
Print the Model Final Result: 王浩 is 王浩
Current score is  346
Test  363
```

```
The Print Sentence is  晓林大哥，你工作要回广州吗？
Stage One: Find the -> Title <- in the paragraph: ['晓林']
The Print Person Name is 范晓林
Print the Model Final Result: 晓林 is 范晓林
Current score is  347
Test  365
```

```
The Print Sentence is  王书记，可以帮忙办理入党事宜吗？
Stage One: Find the -> Title <- in the paragraph: ['王']
The Print Person Name is 王俊成
Print the Model Final Result: 王 is 王俊成
Current score is  381
Test  400
```

2. 后期人名规则匹配张冠李戴-错 4 条

e.g. “李大哥生日礼物你收到了吧？”中“李大哥”原本指的是“李大明”。但是第二个输入是“李鹏飞”，当时是通过“李大哥”拆分为“李”、“大”、“哥”与输入的“李”、“鹏”、“飞”做匹配，此时二者均有“李”， 一旦有一个或多个相同的字，则认为“李大哥”是“李鹏飞”。但是事实上，“李大哥”是“李大明”，导致了张冠李戴的错误。

此类属于 含有称谓的 data + 错误的全称大名 这一类。

e.g. “最近***戴宇航***学习怎么样啊？” + “***戴海生***”（两者均有“*戴*”字，但二者不是同一人）

或者 “***张舒越***最近在忙什么呢？” + “***贾舒越***”（两者均有“舒越”，但二者不是同一人），此时出现问题的原因主要是基于“字”层面的匹配←单独的“字”或“词”并无法体现名称间的语义信息。

问题在于，如何证明部分案例的输入的全称大名不是 NER 命名实体识别后的名称？

解决方案：

（1） 若 NER 识别词⊆全称大名，则 NER 识别词为全称大名的一个称谓;

（2） 若 NER 识别词与全称大名有交集，则：建立一个常用的词典库，如果 NER 命名实体识别后的名称部分与输入的全称大名的交集外的部分（“字”层面）出现在了[字典库+全称大名]中，认为识别的词为全称大名的一个称谓。

例如：[“张”，“玉”，“龙”]（输入的全称大名）与[“小”，“龙”]（NER 命名实体识别）问题，由于[“小”，“龙”] ⊆[“张”，“玉”，“龙”]，则认为[“小”，“龙”]为[“张”，“玉”，“龙”]的一个称谓。

例如：[“戴”，“海”，“生”]（输入的全称大名）与[“戴”，“宇”，“航”]（NER 命名实体识别）问题，[“戴”，“海”，“生”] 与[“戴”，“宇”，“航”]中均有“戴”，则比较[“宇”，“航”]二字是否出现在字典库和[“戴”，“海”，“生”]中，若出现，则认为 NER 命名实体识别提取出的词为全称大名的一个称谓，否则不是。

例如：[“贾”，“舒”，“越”]（输入的全称大名）与[“张”，“舒”，“越”]（NER 命名实体识别）问题，[“贾”，“舒”，“越”] 与[“张”，“舒”，“越”]中均有“舒”与“越”，则比较[“张”]字是否出现在字典库和[“贾”，“舒”，“越”]中，若出现，则认为 NER 命名实体识别提取出的词为全称大名的一个称谓，否则不是。

```
The Print Sentence is  李大哥生日礼物你收到了吧
Stage One: Find the -> Title <- in the paragraph: ['李大哥']
The Print Person Name is 李鹏飞
Print the Model Final Result: 李大哥 is 李鹏飞
Current score is  122
Test  130
```

```
The Print Sentence is  李老师，您最近什么时候有时间呢？
Stage One: Find the -> Title <- in the paragraph: ['李老师']
The Print Person Name is 李鹏飞
Print the Model Final Result: 李老师 is 李鹏飞
Current score is  131
Test  140
```

```
The Print Sentence is  孟宇，我超级崇拜你，北大的高材生！
Stage One: Find the -> Title <- in the paragraph: ['孟宇']
The Print Person Name is 钱开宇
Print the Model Final Result: 孟宇 is 钱开宇
Current score is  131
Test  141
```

已解决

```
The Print Sentence is  乔大哥，你啥时候放假回家呀？
Stage One: Find the -> Title <- in the paragraph: ['乔大哥']
The Print Person Name is 李大明
Print the Model Final Result: 乔大哥 is 李大明
Current score is  282
Test  293
```

3. 规则匹配时候导致错误-错 4 条

之前规则匹配主要是如果识别的是[“郝”，“老师”]这种将一个整体称呼识别拆分为了 2 个，就将 2 者合并为一个，即[“郝”，“老师”]变为[“郝老师”]。(具体有：“老*”、“*老师”、“*老板”、“*处长”、“*局长”、“*将军”、“*司令”、“*主席”、“*先生”、“*女士”、“*师兄”、“*师姐”、“*同学”、“*校长”、“*兄”、“*弟”、“*姐”、“*妹”)。

此时如果一句话中前面有一个称呼，很靠后也有一个称呼的话，合并后反而弄巧成拙。

```
The Print Sentence is  思民，你在哪个老师的实验室干活呢？
Stage One: Find the -> Title <- in the paragraph: ['思民老师']
The Print Person Name is 李思民
Print the Model Final Result: 思民老师 is 李思民
Current score is  67
Test  70
```

```
The Print Sentence is  芳师姐，你有男朋友吗？
Stage One: Find the -> Title <- in the paragraph: ['朋友师姐']
The Print Person Name is 王芳
Print the Model Final Result: There is no match for 王芳
Current score is  76
Test  82
```

```
The Print Sentence is  韬哥，我今天下午和同学一起游游清华校园，就早点走啦！
Stage One: Find the -> Title <- in the paragraph: ['韬哥同学']
The Print Person Name is 代文韬
Print the Model Final Result: 韬哥同学 is 代文韬
Current score is  302
Test  315
```

```
The Print Sentence is  清华大学的校长是谁呢，海生？
Stage One: Find the -> Title <- in the paragraph: ['海生校长']
The Print Person Name is 戴海生
Print the Model Final Result: 海生校长 is 戴海生
Current score is  330
Test  345
```

```
The Print Sentence is  侯龙岳是不是你的小学同学呢？
Stage One - Find the Title: ['侯龙岳同学']
The Input Person Name is 侯欣雨
Stage Two - Print the Model Final Result: 侯龙岳同学 is 侯欣雨
Current score is  537
Test  556
```