查阅资料和仔细思索后，如果要使用神经关系提取，要用bert作为训练和预训练模型，数据标注上还是需要用BIO方法。至于如何结合词性之类的可以提高精确度的NLP方法，可以并行考虑。但是因为要用bert训练，数据集还是需要按照神经训练的需要来标注。我找到了一篇用bert实现关系抽取的博客<https://blog.csdn.net/weixin_42001089/article/details/97657149>。里面一个思路对于之前多谓语的情况有很好的启发：两套训练模型，第一套用来找出有多少种关系（一个句子里会有多个三元组），第二套模型才用来把所有关系都确定下来找出三元组。这样可以找到文章中的所有关系

**之前找的那些NLP方法，词性标注之类的，可以单独作为关系分类的时候的输入，源数据的标注还是只要标注主体客体和关系就可以了。这样套bert也好套。**

**比如主体用 B-SUB I-SUB I-SUB 等表示,客体用 B-OBJ I-OBJ I-OBJ表示，关系单独标注在每句话的右边用一个单字母表示。具体解释看文档**

**预处理数据标签说明**

特殊字符：

[CLS] : Classification 出现在文段的开头，一个文段只有一个[CLS]标记

[SEP] : Separation 每一句话结束有一个该特殊字符，用于分句

BIO:

B: 实体的开头用B-

I: 实体中间的字用I- (最后一个字也是中间)

O：除了B和I以外的都是O

关系只标注这句话，不再句内的字上体现

关系标签：

<定义「D」>：（Definition）

<特征「C」>：（Character）

<目的「G」>：（Goal）

<因果「R」>：（Reason&&Result）

<举例「F」>：（For example）

<问题「Q」>：（Question）

<祈使「S」>：（Should）

<所有关系「H」>：（Have）

<时间于「T」>（time）

<位置于「I」>（in）

标注示例见下面

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

另外附一个用bert实现关系抽取的github项目

https://github.com/yuanxiaosc/Entity-Relation-Extraction

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

关系解释：和之前的文本相同，可以跳过。定义中示例有标注方法示例

**1. 定义**：所有解释性文本，实质上分为解释和归纳两种。解释是后面文段比较长，是前面文段的解释；归纳是前面文段比较长，是后面文段的解释。但这两种都可以属于定义类，为了方便，都划分为“定义”关系。定义类通常谓语中通常是“是”“为”“即”“也就是”，但注意定义类的主语和宾语必须描述的是同等级同一事物。另外标点符号’:’’——’也有“定义关系”【该类关系注意与“特征”关系的区别】

示例：

B-SUB I-SUB I-SUB O O O O O O O O O B-OBJ I-OBJ I-OBJ I-OBJ

石 墨 烯 （Graphene） 是一种 由 单 层 碳 I-OBJ I-OBJ I-OBJ I-OBJ I-OBJ I-OBJ I-OBJ I-OBJ I-OBJ I-OBJ

原 子 组 成 六 角 型 呈 蜂 巢 I-OBJ I-OBJ I-OBJ I-OBJ I-OBJ I-OBJ I-OBJ I-OBJ I-OBJ I-OBJ

晶 格 的 片 状 结 构 的 新 材

I-OBJ O O O O S-OBJ I-OBJ I-OBJ I-OBJ I-OBJ I-OBJ I-OBJ

料 ， 是 只 有 一 个 碳 原 子 厚 度I-OBJ I-OBJ I-OBJ I-OBJ I-OBJ I-OBJ I-OBJ

的 二 维 晶 体 材 料。

提取：[“石墨烯（Graphene）”, “定义”, “一种由单层碳原子组成六角型呈蜂巢晶格的片状结构的新材料，是只有一个碳原子厚度的二维晶体材料”]

**2.特征**：表述主体的特征，通常是形容词的，或者中文中一些隐式表达可以转化成形容词。通常为“是……什么的”“好似”“就像”“……似的”隐式表达比较难判断

示例（常见）：

爱情 是 突如其来的，又是温软迟钝的；是甜美如梦的，又是酸涩难言的

提取：[“爱情”，“特征”，“ 是突如其来的，又是温软迟钝的；是甜美如梦的，又是酸涩难言的”]

示例（隐式）：

冠状病毒粒子 外包着 囊膜，膜表面 分别有 刺突糖蛋白、小包膜糖蛋白和膜糖蛋白

等3种蛋白

【可转化为】：冠状病毒粒子是外包着囊膜的，是膜表面分别有刺突糖蛋白、小包膜糖蛋白和膜糖蛋白等3种蛋白的

提取：[“冠状病毒粒子”，“特征”，“ 外包着囊膜，膜表面分别有刺突糖蛋白、小包膜糖蛋白和膜糖蛋白等3种蛋白”]

**3.目的**：目的是做事情的动机，在于意图层面而不在于是否已经发生，带有主观色彩。谓语通常有“为了”“为的是”“目的是”“确保”“以便”“以此来”“来”。【目的应该与结果区别，目的是主观的意图，结果更侧重发生了】

示例：

为了 减少 航天员 携带这些外在的微生物到太空，在 航天员 进入载人航天器前，

首先要进行检疫。检疫通过后，在起飞前的7到10天，航天员还要进行医学隔

离，确保 尽量减少携带一般微生物登上太空的可能。

提取：

[“在航天员进入载人航天器前，首先要进行检疫”, “目的”, “减少航天员携带这些外在的微生物到太空”]

[“检疫通过后，在起飞前的7到10天，航天员还要进行医学隔离”, “目的”, “尽量减少携带一般微生物登上太空的可能”]

4.因果：表示原因和结果，标注时统一原因是主体，结果是客体，关系为因果。因果类的识别比较简单，有“因为”“所以”“导致”“以至于”“之所以”“之后”。不过有一些隐式的因果需要识别。（已经发生的或者预测未来很可能发生的）

示例（常见）：

因为 池子太深了，所以他不敢一个人游泳

提取：[“池子太深了”，“因果”，“ 他不敢一个人游泳”]

5.举例：分为举实例和同类举例。识别词有“比如”“例如”“如”“就像”

另外有隐式的举例会在句首点明主题之后，后面就用排比来解释

示例（常见）：

这世上 有 很多种毒， 比如 蜈蚣毒、蛇毒和孤毒。

提取：[“这世上有很多种毒”，“举例”，“ 蜈蚣毒、蛇毒和孤毒”]

示例（隐式）：

这世上 也 有很多种死法，吊死、闷死和伤心死。

提取：[“这世上也有很多种死法”，“举例”，“ 吊死、闷死和伤心死”]

6.时间：比较容易识别“在”“的时候”“于”，但是有时候需要拼接，因为时间状语可能出现在任何位置，这一点于地点状语一样

示例（常见）：

1949年，中华人民共和国成立

提取：[“中华人民共和国成立”，“时间”，“ 1949年”]

示例（拼接型）：

中华人民共和国于1949年成立

提取：[“中华人民共和国成立”，“时间”，“ 1949年”]

7.地点：时间和地点的识别词比较像，具体分析

示例（常见）：

在天安门上，毛主席发表讲话

提取：[“毛主席发表讲话”，“地点”，“天安门上”]

示例（拼接型）：

毛主席在天安门上发表讲话

提取：[“毛主席发表讲话”，“地点”，“天安门上”]

8.问题：由问题对象和问题内容构成。“为什么”“什么”“怎么”“何”

示例：

冠状病毒是什么？

提取：[“冠状病毒”，“问题”，“是什么”]

9.祈使：表示未来最应该/不应该怎么做，未来发展的趋向最符合推理的正确道路或者最符合推理的错误道路。通常是祈使句，用来表示观点，明确内容。“应该”“要”“就要”“不应该”“不能”。由祈使对象，祈使内容组成。会遇到省略祈使对象的情况，默认“读者”

示例：我们应该从身边的小事做起

提取：[“我们”，“祈使”，“应该从小事做起”]

>

10.所有关系：所有权，控制权，使用权

示例：

我有一把吉他