# 机器翻译作业-关系抽取

近年来信息爆炸式增长，尤其是杂乱无序的非结构化、半结构化信息的增长，使得传统的基于关键词匹配技术和PageRank等网页排序技术的搜索引擎虽然能够在一定程度上满足用户信息获取的需求，但往往需要一定的人工筛选才能找到自己想要获取的信息。因此，信息抽取技术的发展体现了人们对于信息获取的快捷性、准确性和方便性的要求。

信息抽取技术是NLP中的重要任务，在机器阅读理解，对话生成，推荐系统中都有广泛的应用。也是知识图谱体系结构的初始环节也是最重要的环节。信息抽取的质量好坏直接影响下流任务的质量。 而信息抽取技术也一直是自然语言处理的研究热点，受到了学术界和工业界的广泛关注。

信息抽取包含实体识别，实体关系抽取，事件抽取三个部分，实体关系抽取是信息抽取的重要的研究课题，也是核心任务。实体关系抽取的任务是从一段文本或者句子里面抽取该句子提及的两个实体之间的语义关系。从近些年的论文来看，大家对实体关系抽取分成了两种方式来做，一种是直接从非结构化文本中抽取三元组。另一种是当做关系分类，句子中的实体对是事先标记好的，即是在已经做好命名实体识别和实体标注的基础上再来确定实体对之间的语义关系。而实体识别则是信息抽取的另一个重要的研究课题了。关系抽取是其重要的子任务之一，主要目的是从文本中识别实体并抽取实体之间的语义关系。  
举例：  
{  
“text”: “《邪少兵王》是冰火未央写的网络小说连载于旗峰天下”,  
“spo\_list”:  
[{“predicate”: “作者”,  
“object\_type”: {"@value": “人物”},  
“subject\_type”: “图书作品”,  
“object”: {"@value": “冰火未央”},  
“subject”: “邪少兵王”}]  
}  
 上述是[百度2020年语言与智能技术竞赛：关系抽取任务](https://aistudio.baidu.com/aistudio/competition/detail/31)中的一个例子，通过给定的文本，识别出其中的主实体为邪少兵王，客实体为冰火未央，主实体和客实体的关系为作者，s是主体，o是客体，p是关系，也叫做spo三元组。

关于关系抽取的研究可以归纳为：主流的研究多为端到端模型（End-to-End Models）或者是线性模型（Pipeline Models），主要采用的为深度学习方法，所用多为卷积神经网络（Convolutional Neural Networks, CNN）或者是循环神经网络（Recurrent Neural Network, RNN），图神经网络(Graph Neural Network, GNN)在一部分论文中也得到了应用；2018年，Google AI提出基于Transformer的BERT预训练模型后，在多项自然语言处理（Natural Language Processing, NLP）任务上的表现刷新了记录；随后，各种基于Transformer预训练模型如雨后春笋般出现；现在，NLP任务多数都采用Transformer结构作为自己模型的基础编码部分。近年来，随着深度学习的发展，各种深度神经网络模型层出不穷，模板匹配的方式早已经被淘汰，选取复杂的特征或优化核函数等传统的机器学习方法也逐渐被研究者们放弃了，转向使用深度学习的方法来研究关系抽取。特别是2018年Google公司发布基于Transformer的BERT预训练模型，其在多项自然语言处理（Natural Language Processing,NLP）任务上的表现刷新了记录。随后各种基于Transformer预训练模型如雨后春笋般出现，特别是参数上千亿的GPT-3, 甚至直接能够用来生成作文，在很多NLP任务上的性能直接超过人类。

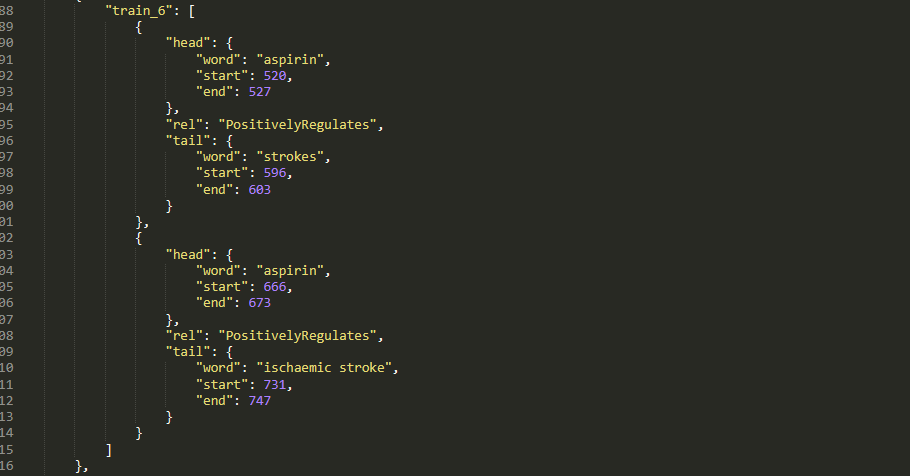
使用BERT这样的预训练语言模型对输入本文建模，得到的文本表征几乎能包含整个文本的所有语义信息，但是并不是说语言学的知识已经失去了实用性。事实上，在深度神经网络中引入语言学的信息，能够让模型在特定的NLP任务上性能更好。所以，寻找一种可行的方法来引入语言学信息，来辅助关系抽取任务仍然存在巨大的潜力。使用BERT等基于Transformer预训练模型进行关系抽取研究已经成为未来此领域的潮流之一。在我的研究中，也将采用此方法，再结合文本的语言学信息，构建一种端到端模型(End-to-End Models)来实现从原始句子中抽取实体之间的关系。

Casrel是发布在2020acl上的一篇关于关系抽取的论文，这篇论文也是百度关系抽取比赛的一个原型。它的主要思路是先通过一个编码器来预测主体，为了解决实体嵌套的问题，它采用了基于span的方式预测。之后基于每一个预测的主体去结合关系类型预测客体，对每一个关系类型独立设计一个客体识别的网络，也是基于span的方式，这样就很好的解决了实体嵌套和关系重叠的问题。

作者开放了源代码，但是它的源码有者很大的局限性：（1）在编码寻找主客体所在的位置时，直接采用的是字符串find方法，这样的话针对短序列可以，因为在短序列中主客体一般只出现一次，但是在长文本中往往会出现主客体出现多次的情况，这种情况就不能被很好的解决。（2）输入长度的限制，由于模型架构采用的是bert，输入会经过BPE，这样的话原始的索引可能出现偏移，例如单词unhappy，经过bpe之后会分成un和happy两个token。为了对模型的输出进行恢复，在输入的时候对原始token之间会加一个unk，这样可以在恢复的时候更方便。但是这样的话会增大模型输入的长度，尤其是bert一般限制最大输入长度为512。

**我的工作**：（1）去掉增加的unk操作，采用一种mask掩码的方式去恢复最后的输出，增大了模型对长文本序列的适应能力。（2）采用逐token去bpe，解决文本中可能出现多次主客体的问题。（3）对原始开源代码进行重构，形成自己的代码风格。（4）在改进的基础上参加“链想家英文关系抽取”比赛，改进版的代码比原始高了接近10个点。

下面我将以“链想家关系抽取”比赛为原型展开叙述。比赛提供了一些关于convid-19的论文摘要和引言部分，属于长序列文本，每个文本中存在着多个关系三元组，并且每个实体可能组成多个关系。例如SARS-COV-2与convid-19存在着cause关系，而convid-19与fever也存在着cause关系。并且原始的英文文本比较不规范，标点符号都是和实体连在一起，并且对于某个词，它的标注实体可能只是这个词的一部分，这就大大的增大了难度。



**解决思路**：

1. 数据预处理。这部分工作主要是将原始text中和标注实体连在一起的特殊符号拆成以空格划分的单个token，这样对于之后模型输入的索引问题能够很好的解决。这部分代码很多，我将放在github上进行展示。这部分的难点是将特殊字符转成独立的token之后原始的实体index偏移问题的解决。
2. 模型输入。划分数据为验证集和测试集，为了方便模型处理，提前将数据处理成pk格式文件。

Safety and efficacy of intravenous bimagrumab in inclusion body myositis (RESILIENT): a randomised, double-blind, placebo-controlled phase 2b trial Bimagrumab showed a good safety profile, relative to placebo, in individuals with inclusion body myositis but did not improve 6MWD. The strengths of our study are that, to the best of our knowledge, it is the largest randomised controlled trial done in people with inclusion body myositis, and it provides important natural history data over 12 months.

['PositivelyRegulates']

[18]

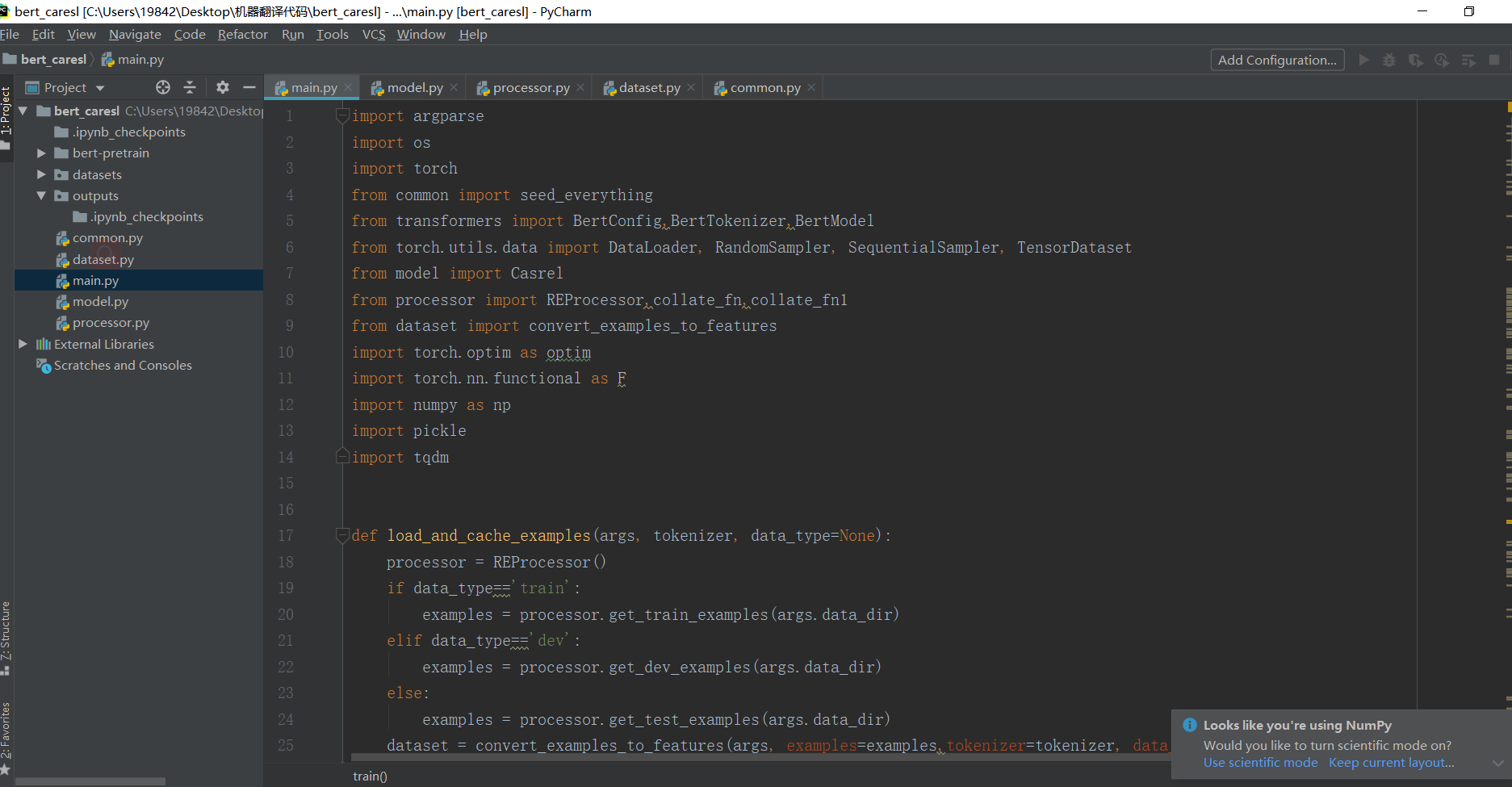
[30]

[18]

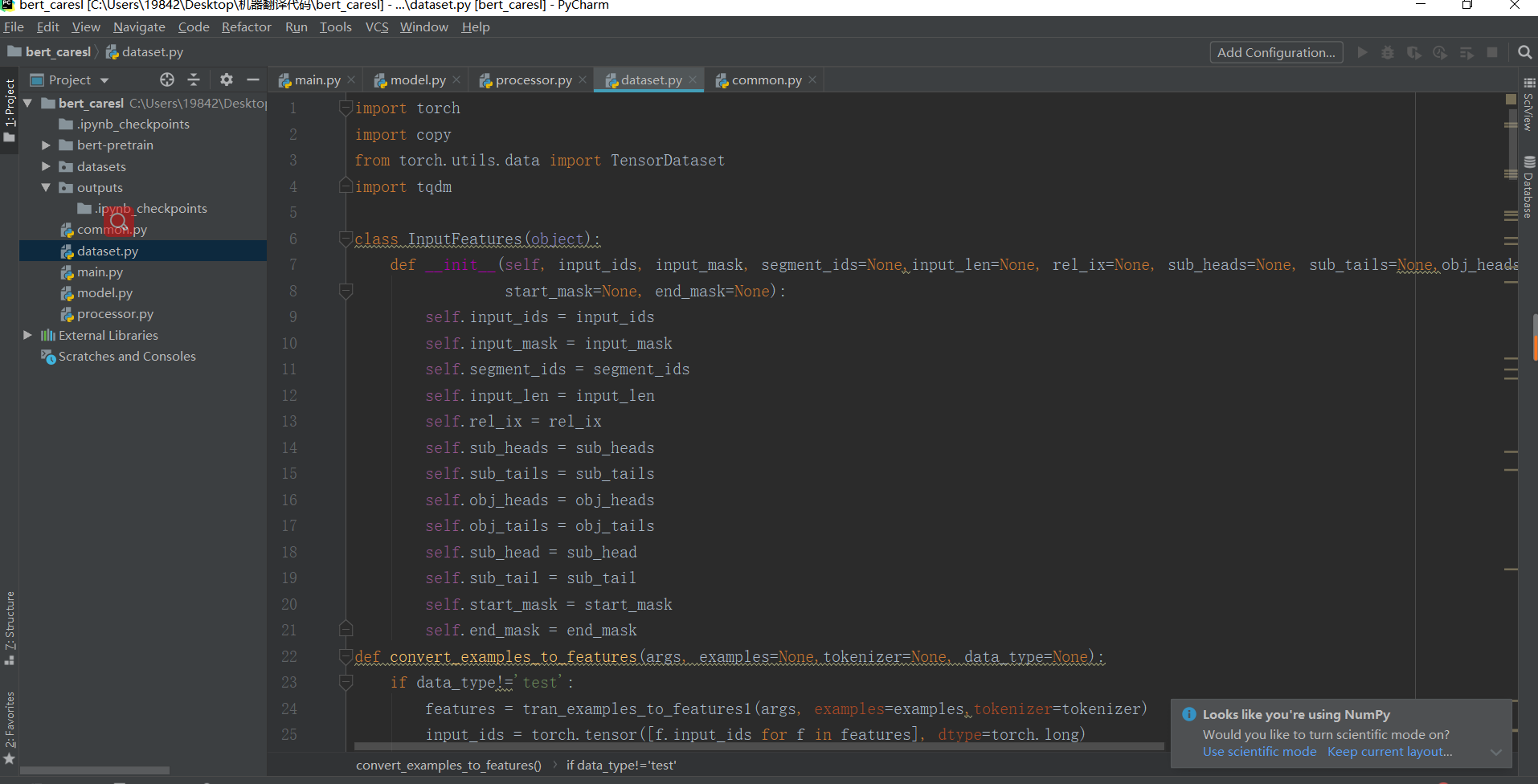
[32]

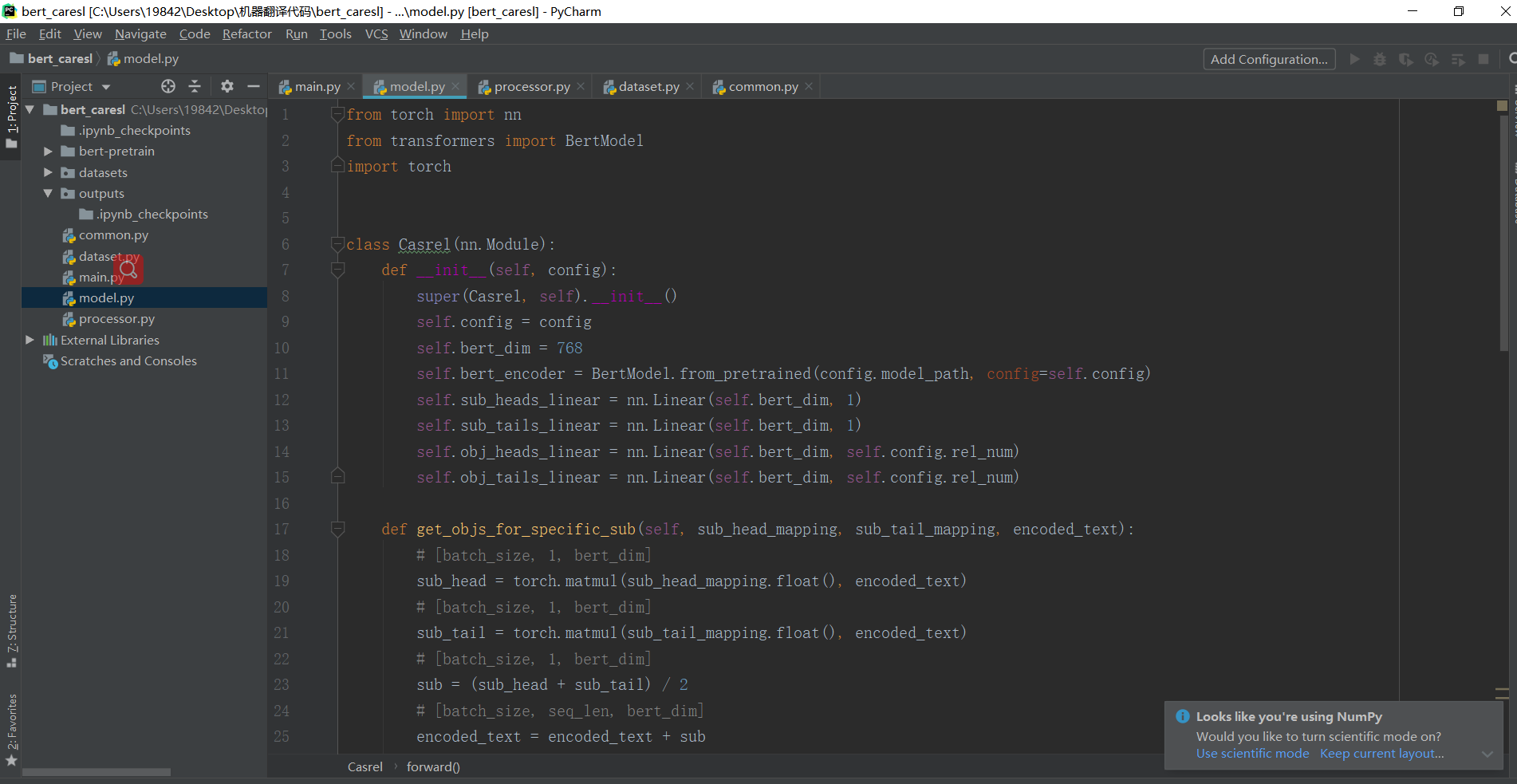
上面是一条数据的输入格式，'PositivelyRegulates'表示该text中的一个关系类型，两个18表示主替的索引，在该句中主体就为’Bimagrumab’，30和32表示客体的索引，在该句中客体就为’inclusion body myositis’。

1. 模型框架流程。



（1）Main.py。程序的执行入口，包括定义的一些模型参数，包括学习率，batchsize、epoch、学习率策略等等，还包括训练和预测的详细流程。

（2）dataset.py。主要包括对输入数据的一些处理，处理成bert的输入格式。



（3）Model.py。包括定义的的模型框架，由bert和若干个linear层组成，对于start，linear预测0或者1，表示该个token是否是实体的开始。同样的对于end，linear预测0或1，表示该个token是否是实体的结束。最后通过一个掩码操作来恢复预测的的主体和客体。

最终结果：前期我们采用使用carsel来进行联合关系抽取，使用原始论文的github源码来处理这个任务只能达到25左右的分数，但是采用这种改进之后的框架能达到35左右的分数，足足高了十个点。但是由于此任务的特殊性采用联合关系的方法效果不是很理想，后期我们采用了基于管道的方法使效果大大改进，最终取得了a榜第一b榜第三的成绩。



总结：

关系抽取作为智能问答系统、搜索引擎和推荐系统等实际应用中的重要上游组成部分，如何更加准确地抽取实用信息是学术界和工业界的一个研究热点。同时，关系抽取技术在海量信息处理，知识库自动构建和搜索引擎等领域具有重要意义：通过关系抽取技术，从无结构的自然语言文本中抽取出格式统一的关系数据，有助于计算机快速处理大规模文本数据，提高处理效率；通过抽取实体之间的语义关系，能够建立多个实体之间广泛的信息关联，有助于建立领域本体，帮助知识图谱的构建；通过深入挖掘和分析自然语言文本中的语义关系信息，能够进一步理解和匹配用户的查询意图，从而为用户提供更精准的搜索服务。通过参加这次比赛不但让我提前熟悉了实验室的研究方向，而且加深了我对关系抽取的理解和认识，提高了自己的编码能力。