**1.4 CRF层能从训练数据中获得约束性的规则**

CRF层可以为最后预测的标签添加一些约束来保证预测的标签是合法的。在训练数据训练过程中，这些约束可以通过CRF层自动学习到。  
这些约束可以是：  
I：句子中第一个词总是以标签“B-“ 或 “O”开始，而不是“I-”  
II：标签“B-label1 I-label2 I-label3 I-…”,label1, label2, label3应该属于同一类实体。例如，“B-Person I-Person” 是合法的序列, 但是“B-Person I-Organization” 是非法标签序列.  
III：标签序列“O I-label” is 非法的.实体标签的首个标签应该是 “B-“ ，而非 “I-“, 换句话说,有效的标签序列应该是“O B-label”。  
有了这些约束，标签序列预测中非法序列出现的概率将会大大降低。

这样的话，那些词性分析的时候，我是不是变成 B-N I-N更合适，这样能够保证连贯性

 在英文文本中单词为最小单位，且每个单词均有一一对应的词性信息，所以无需过多的预处理。但对中文文本，BERT模型的最小输入单位为单个汉字， 但是词性信息是根据词语进行标注的，为了满足BERT的要求，我们需要对数据进行预处理，将原文本拆分成一系列的汉字，并对每个汉字进行词性标注。这种分词处理有多种体系，这里使用 ”BIO“，其中 “B” 表示该汉字是词汇开始字符，同时也可以表示单字词；“I” 表示该汉字是词汇的中间字符；“O” 表示该汉字不在词汇当中。”O” 在词性标注任务当中不会出现，但是在命名实体标注中有意义。经过预处理后的数据如下图所示。此外根据BERT模型的要求，需要预先设定最大序列长度(max\_seq\_length)，根据此参数对序列进行padding。

[ Preprocessed data example ]





