**对抗训练多头选择的实体识别和关系抽取的联合模型**

<https://arxiv.org/pdf/1808.06876.pdf>

https://arxiv.org/pdf/1804.07847.pdf

实体识别和关系抽取的目标是从非结构化的文本中发现实体mention的关系结构，它对知识库的构建使用和问答等任务都很重要，它是信息抽取的核心问题。

现有的方法整体来说有两种：把实体识别和关系抽取看作两个分开的任务的pipline模型、联合模型，它们都存在着缺陷。

现有模型的不足：

pipline模型：错误会在不同组件之间传播，造成错误的累积；关系抽取只是利用了NER任务的结果，造成一个任务中可能对其他任务有用的信息没有被用到**。**

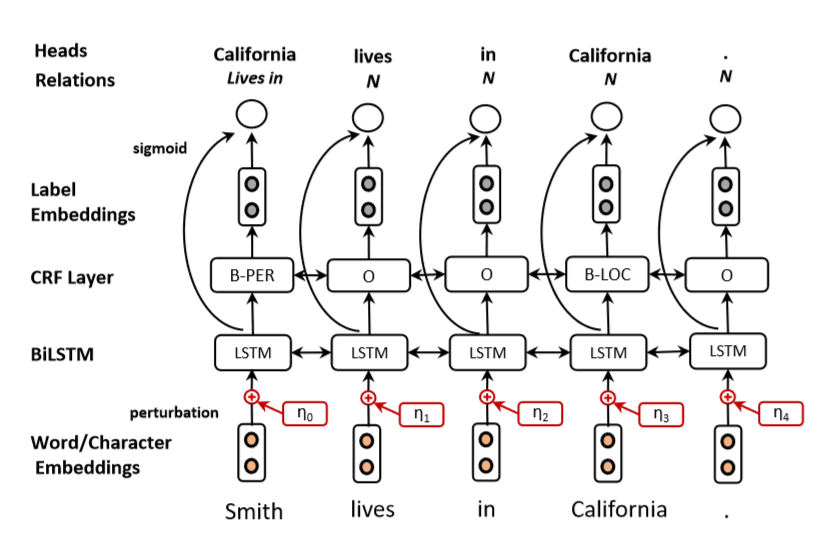
当前的联合模型：有的需要人工特征，虽然利用神经网络已经可以不人工选特征了，但是会依赖于很多NLP工具，如pos标注、依存分析；单个实体的多个关系抽取还存在问题。

本文要介绍的联合模型能一块执行实体识别和关系抽取两个任务，同时能够解决多关系问题，并且不依赖其他的NLP工具，不需要人工设置的特征。

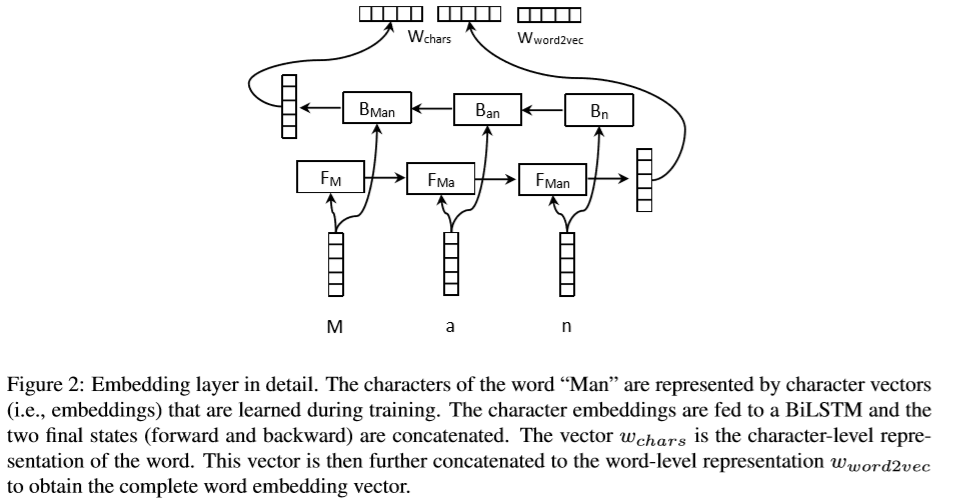
1）将原来的single head selection拓展为预测多个头；

（2）对于头和关系的决策是一块完成的，而不是先预测头，然后在下一步用关系分类器预测关系；

（3）对抗训练；



特征嵌入层：



输入句子（词序列），输出词的embedding序列。

在词级别的向量上加入了字符级的信息，这样的embedding可以捕捉如前缀后缀这样的形态特征，这样的形态学特征在英语、德语这类语言的实体识别任务中效果有显著的提升。

先用skip-gram word2vec模型预训练得到的词向量表把每个词映射为一个词向量，然后把每个词中字母用一个向量表示，把一个词中所包含的字母的向量送入BiLSTM中，把前向和后向两个最终状态和词向量进行拼接，得到词的embedding。

**Bidirectional LSTM encoding layer**

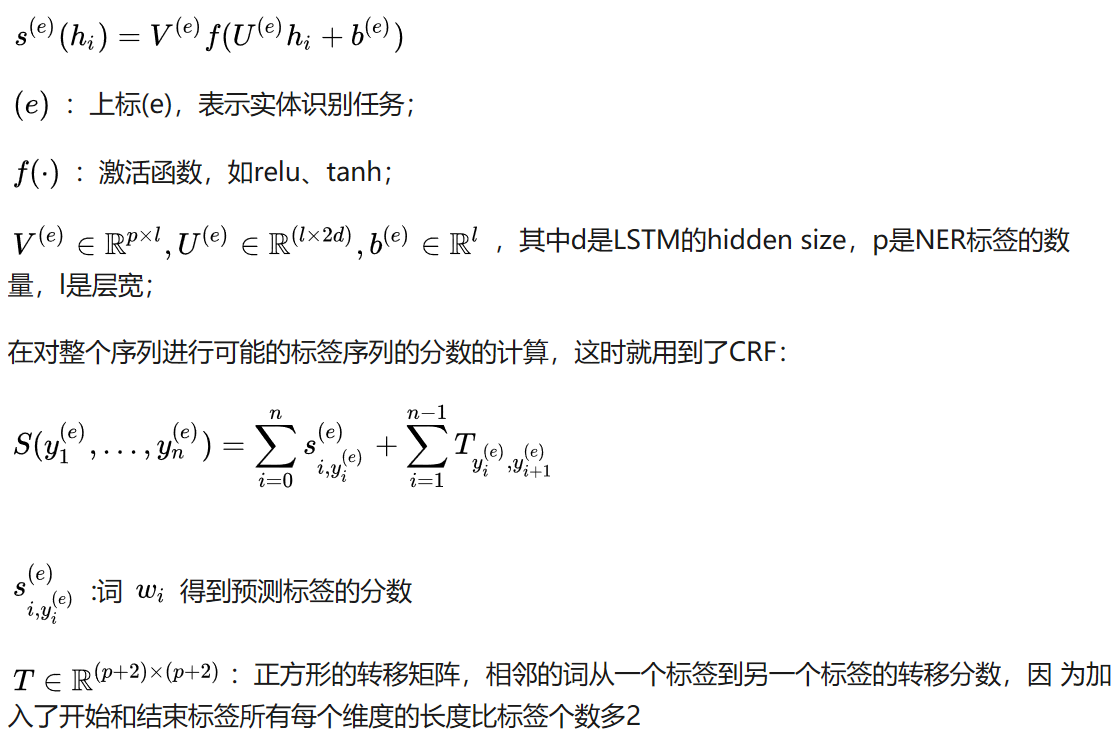
经典的BiLSTM模型，把句子中所包含的词的embedding输入其中，然后将前向和后向每个对应位置的hidden state拼接起来得到新的编码序列：

## 

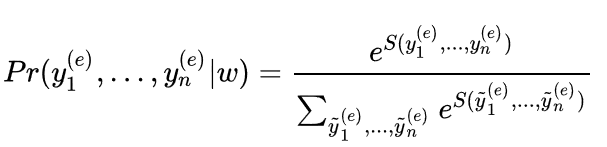
# Named entity recognition

# 采用BIO标注策略，使用CRF引入标签之间的依赖关系。

# 先计算每个词得到不同的标签的分数：



最后计算句子的标签序列概率：



在预测的时候用Viterbi算法得到分数最高的序列标签。

在进行命名实体识别的时候通过最小化交叉熵损失来达到优化网络参数和CRF的目的，在测试的时候用Viterbi算法得到分数最高的标签序列。

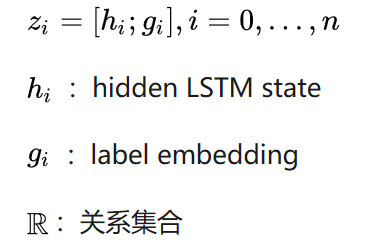
Relation extraction as multi-head selection

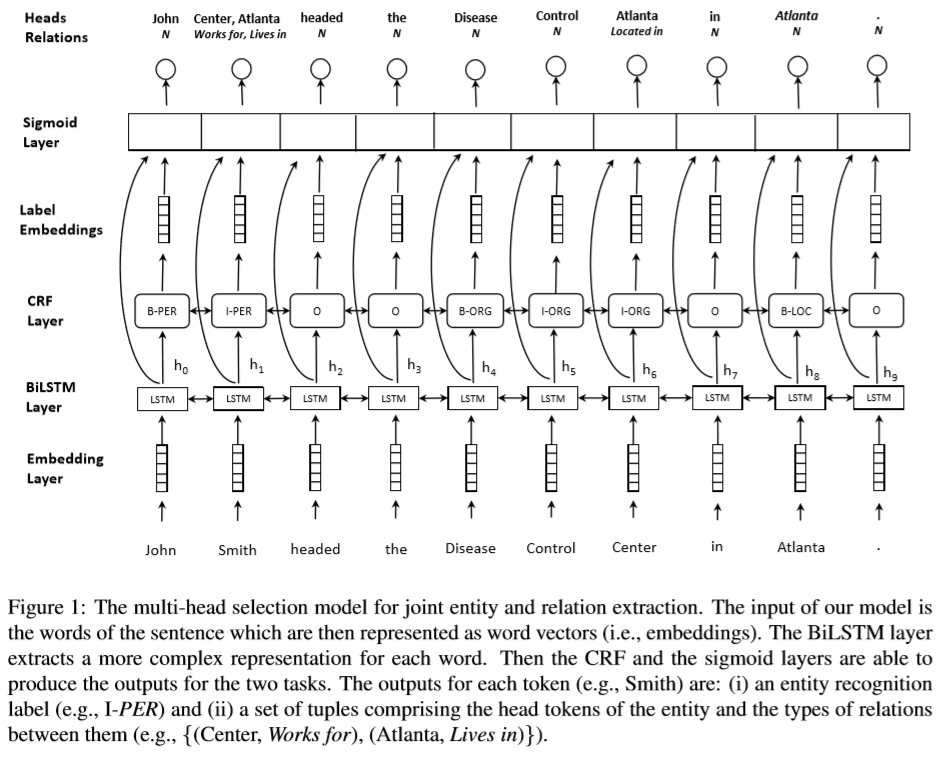
与以前的标准的对依存分析的head selection不同，有了两大创新点：

（1）将原来的single head selection拓展为预测多个头（head就是关系后面的那个词:altanla也就是样本最后一列那个数值，smith live\_in altanla）\*

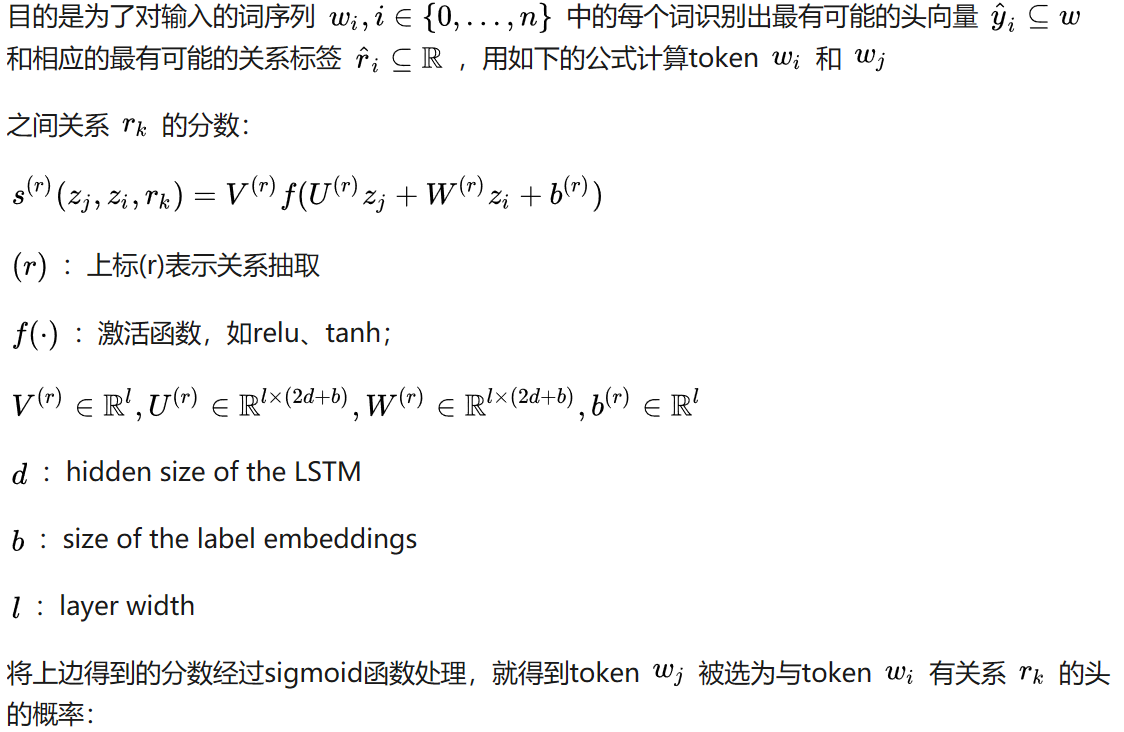
（2）对于头和关系的决策是一块完成的，而不是先预测头，然后在下一步用关系分类器预测关系

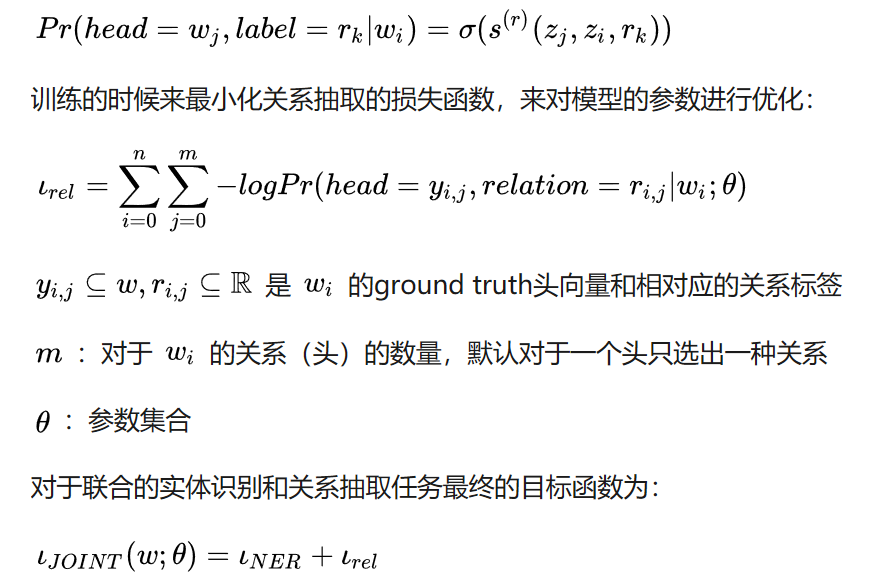
输入为编码BiLSTM的hidden state和实体标签embedding的拼接和关系集合 ，训练的时候标签选用gold entity tags，测试的时候用预测得到的实体标签：





**标签策略**：CRF层的输出是采用BIO标注策略的实体识别结果，heads relations层只有在与其他的实体有关系的实体的尾单词才会给出对应的实体的尾单词和关系；而在与其他实体没有关系的实体和不是实体的单词上，heads为原单词，关系为N。





## Adversarial training(AT)

Goodfellow等人在图像识别中提出的对抗训练可以使得分类器对于有扰动的输入能有更强的鲁棒性，在NLP中也有许多变体被提出用于不同的任务中，如：文本分类、关系抽取等，AT被认为是一种规则的方法，但它不像word droup那样引入随机的噪声，AT生成的扰动是容易被模型错误分类的例子的变形。

