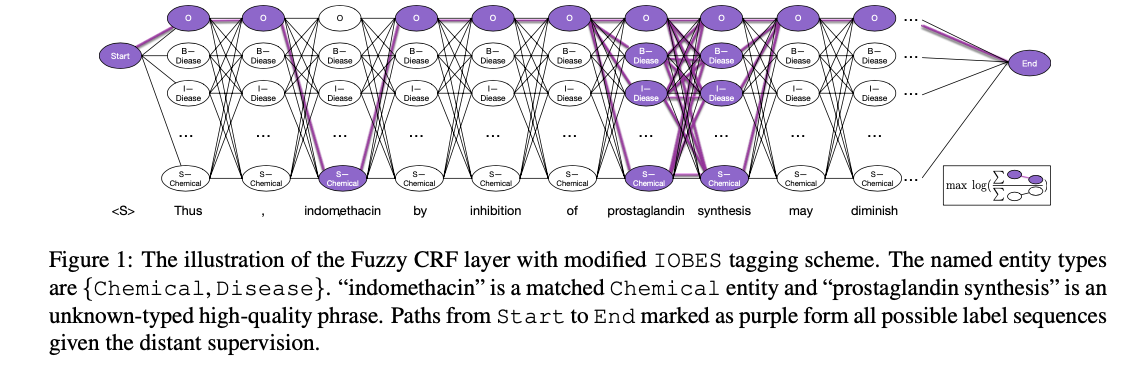
论文阅读笔记一

1. 贡献
2. 提出了一种经过修订的模糊CRF层，该层可以处理具有多个可能标记的标记。
3. 提出了一种神经模型AutoNER和一个新的Tie or Break方案，用Tie or Break来替代IOBES解决远程监督带来的噪声问题。
4. 方法
5. 改进的IOBES标记方案的模糊CRF层



实体类型为{Chemical, Disease}，indomethacin是Chemical实体，它将被标记为S-Chemical；而prostaglandin synthesis是未知类型的高质量短语，它可能被标注为{I, B, E, S}-Disease，O共五种可能。其他词均被标记为O。从Start到end的紫色序列路径是来自远程监督的所有可能的标签序列。

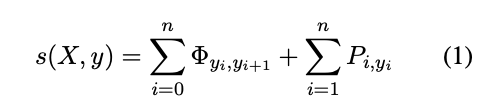
1. 改进的IOBES标注方案

根据三个分词类定义标签。

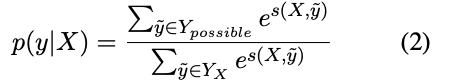
1. 对于标记为一种或多种类型的词，根据词在它匹配实体中的位置用{I,O, B, E, S}中的一种进行标记。
2. 对于未知类型的词，所有五个标签{I,O, B, E, S}都是可能的，同时每个词只有两种类型(Chemical和Disease)，故每个未知类型的词共有9(2×4+1)种可能的标签。
3. 若标记为非实体，则用O表示。
4. Fuzzy-LSTM-CRF

模型在LSTM-CRF基础上进行了修改，以便支持修改后的IOBES标签。

给定单词序列（X1，X2，...，Xn），模型首先对单词进行单词级别的BiLSTM两个方向的编码，得到单词的上下文表示。然后模型为每个输出标签做出独立的标记决策，模型估计单词Xi的标签为Yi的得分为Pi，yi，预测序列的得分定义为：

其中，Φ是从标签i转移到标签i+1的概率，它是一个(k+2)×(k+2)的矩阵，其中k是不同标签的数量。使用两个附加标签start和end分别表示序列的开始和结尾，这两个附加标签仅在CRF层使用。

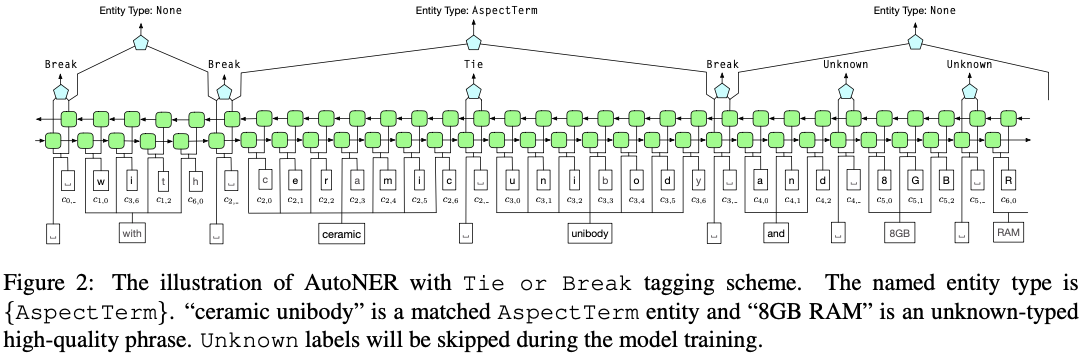
改进的IOBES标注方案中一个实体油腻多种有效实体，故一个句子对应多种有效的标签序列。故将常规CRF扩展为模糊CRF，通过枚举IOBES标签和所有匹配的实体类型来最大化所有可能标签序列的总概率。最大化目标为：

其中，Yx表示X对应的所有可能的标签序列，Ypossible包括所有可能的标签序列。

当所有的标签和类型都是已知的而且都唯一的时候，模糊CRF和常规CRF是等效的。

训练阶段最大化最大似然函数，解码阶段使用viterbi算法。

1. AutoNER with “Tie or Break”



对于带有Tie or Break方案的系统，实体类型为{AspectTerm}，ceramic unibody被匹配为AspectTerm，而8GB RAM是一个未知的高质量短语Unknown，模型训练阶段会跳过Unknown标签。

1. Tie or Break 标注方案

对于两个相邻的词，当两个词匹配为同一个实体时，标注为Tie；

如果至少一个单词属于未知类型高质量短语，则标注为Unknown；

其他情况则标注为Break。

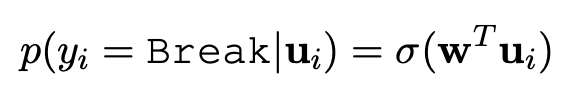
根据Figure2，ceramic unibody被识别为AspectTerm，这两个词用Tie连接，而8GB RAM为高质量短语，故用Unknown标注。

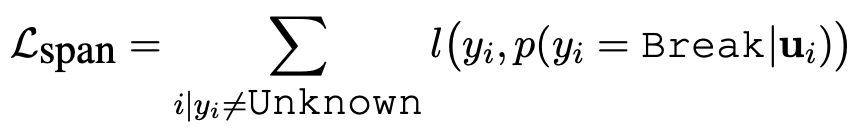
该方案可以更好地利用字典中的知识。

1. AutoNER

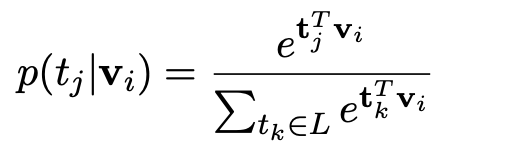
在Tie or Break方案中，实体跨度和实体类型被编码为两个部分，故将实体跨度识别和实体类型预测分为两步。

对于实体跨度检测，作者构建了一个二进制分类器以区分Tie和Break，跳过Unknown标志，具体来说：

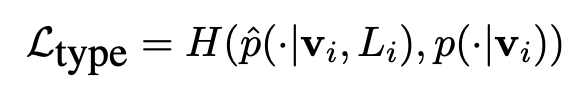
其中，yi表示第i个词和它之前的词之间的标签，σ是激活函数，w是激活函数的参数，实体跨度的损失函数被定义为：

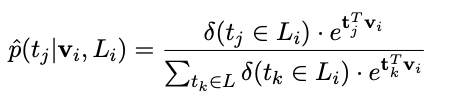
其中l表示logistic回到

得到实体跨度之后需要进行实体的预测：

其中tj是实体类型，L是所有实体类型的集合。

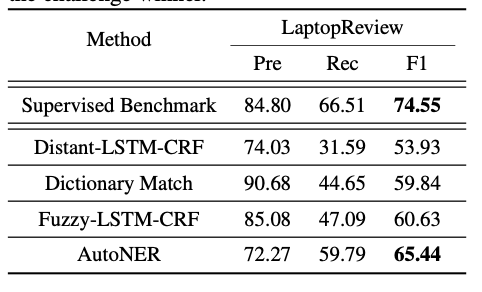
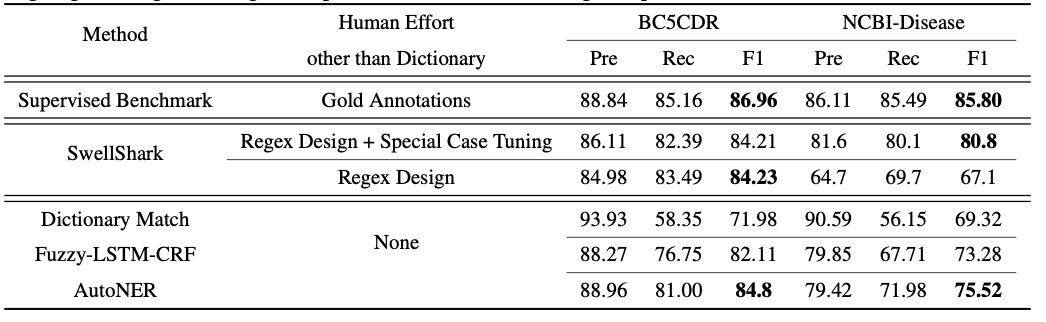
因为一个跨度可被标记为多种类型，故将第i个实体跨度后向类型标记为Li，修改交叉熵损失：



H(p,q)是p和q之间的交叉熵，δ是在远程监督中检查第i个候选跨度是否被标注为实体tj的布尔函数。

AutoNER没有用CRF和viterbi解码。

1. Unknown实体不是其他类型，而是实验无法确定的实体。
2. 实验结果及分析



作者使用AutoNER 和标记数据进行实验，所得结果差异很大。

除了词典数据分布对结果的影响，还用考虑标注方法的有效性以及Tie or Break对模型性能的影响。

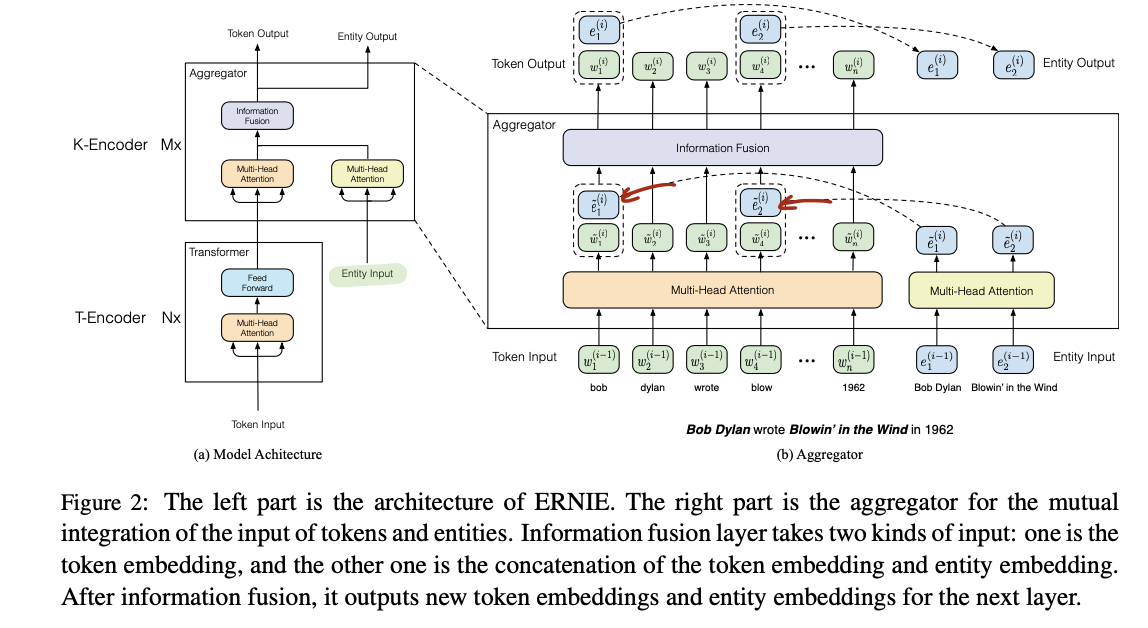
论文阅读笔记二

1. 贡献
2. 作者提出了一种将大量原始语料和图谱中的结构化知识进行编码来训练语言的向量表示的模型ERNIE。
3. 作者定义了一种新的预训练目标，就是随机掩盖输入中某些实体，让模型来预测单词对应的实体类型。
4. 方法
5. 符号

给定句子被表示为{w1, …,wn}，其中n表示句子长度，实体序列表示为{e1, …, em}，其中m表示实体的个数，大多数情况下m并不等于n，

将整个词汇表表示为，实体列表表示为，如果一个词w∈，且有其对应的实体类型e∈，则他们的对齐关系为f(w)=e。

1. 模型结构

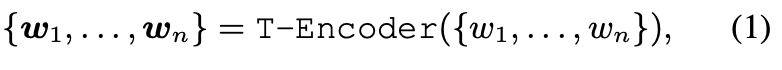


ERNIE模型包括两个模块，图二左下是文本编码器T-encoder，该模块用于获取输入的词法和句法信息；左上是知识编码器K-Encoder，该模块用于整合和提取结构化知识信息，将这些信息集成到文本信息中，以便将单词和实体的异构信息表示为统一的特征信息。

T-encoder层数为N，K-Encoder层数为M。

1. T-Encoder

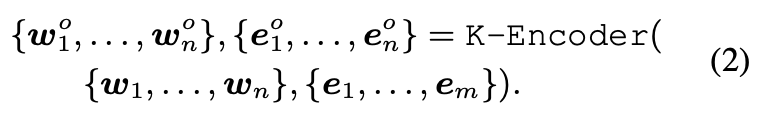
文本编码器将每个词的词嵌入、段嵌入和位置嵌入相加，然后计算词汇和语义特征{w1,…,wn}



其中T-Encoder(.)是一个多层双向transformer编码器，该编码器来自于论文attention is all you need。

1. K-Encoder

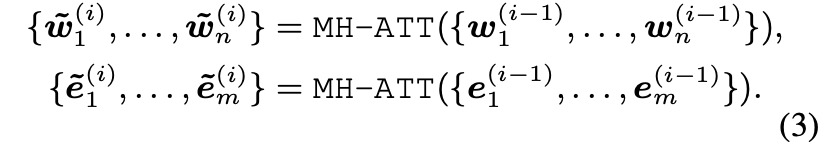
然后ERNIE采用知识编码器K-Encoder将知识信息编码到语言表示中。具体来说，用transE训练单词得到的词嵌入{e1,…em}来表示实体，将{w1,…,wn}和{e1,…em}送到K-Encoder中进行信息融合，最后得到输出的词嵌入：



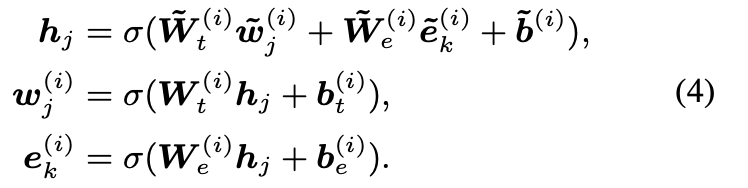
其中w和e将作为特征应用于特定任务。

K-Encoder部分由堆叠的聚合器组成，这些聚合器被用于编码单词和实体，并且可以融合异构特征。

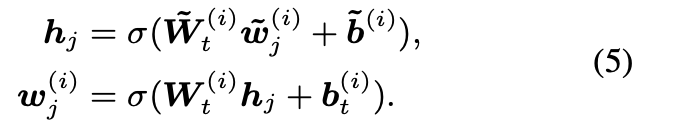
对于第i个聚合器，输入单词{w1,…,wn}和实体{e1,…em}嵌入，这两种词嵌入将被分别送到两个多头自关注层(MH-ATTs)中：

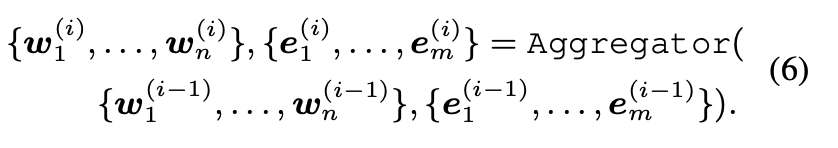


然后该聚合器采用一个信息融合层，将单词和实体序列进行融合，计算每个单词和实体的输出嵌入，对于单词wj对应的实体ek，信息融合过程如下：



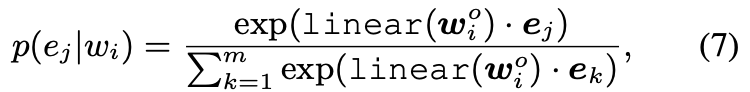
其中hj是集成单词和实体信息的内部隐藏状态，σ(.)是非线性激活函数，通常是GELU函数。对于那些没有对于实体的单词，信息融合层会计算单词的词嵌入而不进行融合：



上述的聚合层可表示为

1. 预训练任务

该任务随机掩盖已经对齐的单词-实体，然后要求系统按照对齐的方式来预测实体。因为该任务与去噪自动编码器，故将处理实体的过程命名为denoising entity token auto-encoder(dEA)。类似对于给定的单词序列{w1,…,wn}和对应的实体{e1,…em}，定义单词的实体分布：

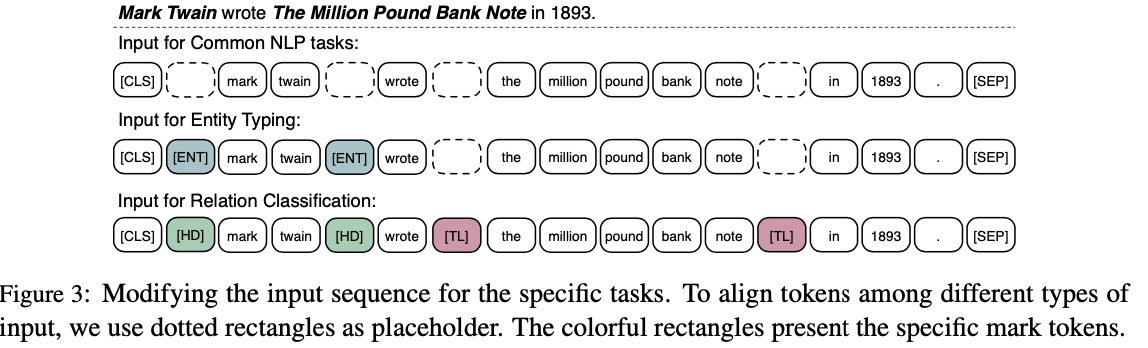


其中linear(.)是线性层，等式(7)将被用于计算交叉熵损失函数。

单词和实体对齐可能会有错误，对 dEA执行以下操作：1) 在5%的时间中，给定单词-实体对，用随机的一个实体来取代当前实体，该操作用于减少模型的预测误差；2) 在15%的时间直接掩盖单词-实体对，用于纠正系统未提取现有对齐的实体的错误；3) 剩余时间中，保持内容不变。

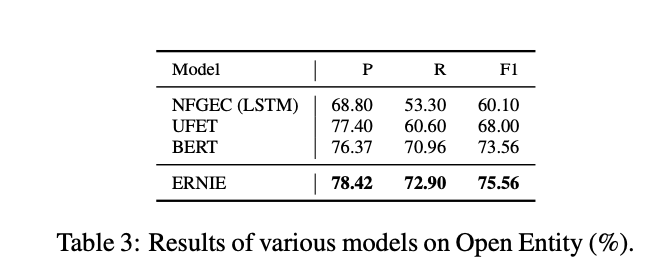
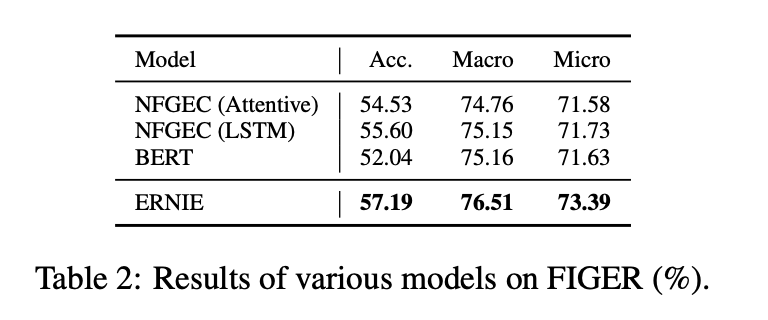
跟BERT类似，ERNIE也采用了MLM和NSP两个预训练任务，使得模型能够从文本的标记中获取词汇和句法信息，具体信息参考BERT原文。总的预训练损失是dEA、MLM、NSP损失的总和。

1. 任务微调



将第一个词的词嵌入作为微调的起始标志，对于关系分类任务，作者通过添加两个标记突出显示实体来修改输入，这些多余的标记类似于位置嵌入，然后作者使用[CLS]标签进行分类。并且设计了头实体[HD]和尾实体[TL]标记。对于实体识别任务，可将其看作简化版的关系分类任务，为输入增加[ENT]标记，从而让ERNIE组合上下文信息和实体信息进行训练。

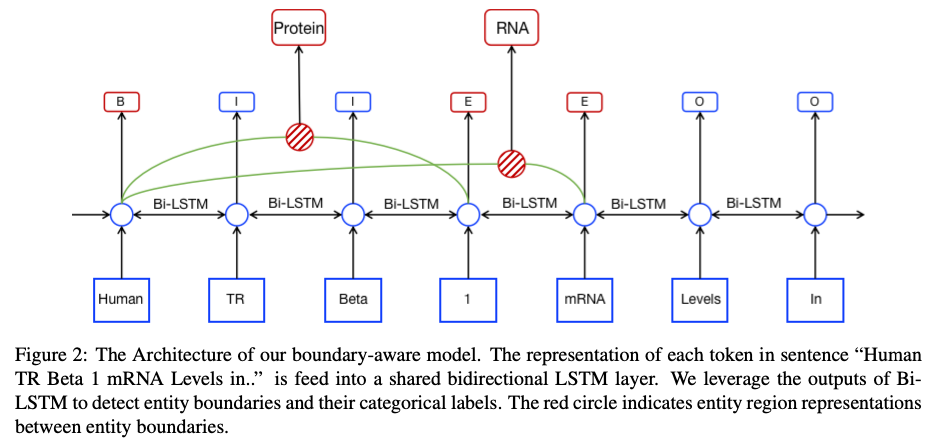
1. 实验结果



ERNIE 通过在输入中注入知识图谱的结构化的信息，有效减少FIGER的噪声标签。

论文阅读笔记三

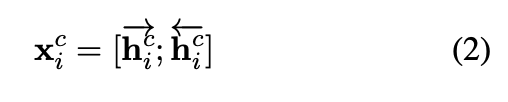
1. 贡献
2. 提出了一种边界感知神经网络用于识别嵌套实体
3. 利用实体边界预测实体类别。
4. 使用序列标记模型检测边界精确定位实体，引入多任务学习来获取实体边界与事件标签之间的依赖关系。
5. 方法
6. 模型结构

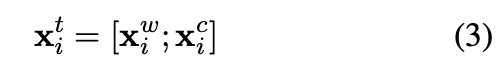


1. 分词的向量表示

对于一个含有n个词的句子{t1,t2, …, tn}，经过在单词向量表中查找、BiLSTM训练和拼接得到单词的向量表示：

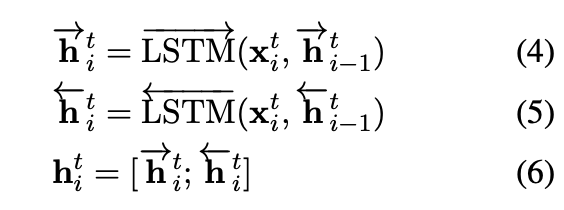






1. 共享特征提取器

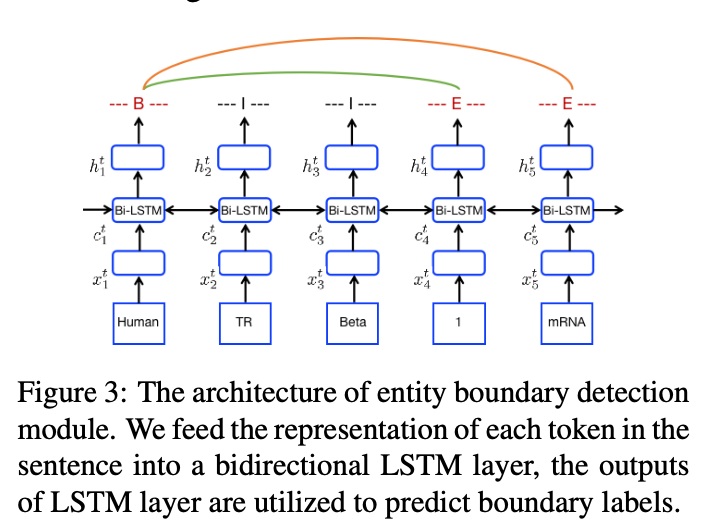
得到单词的向量表示之后，用BiLSTM抽取上下文特征：



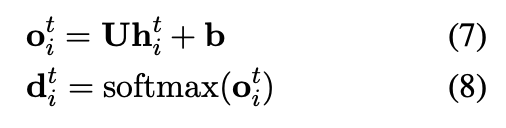
其中x来自前面的上下文表示，h是隐藏层状态向量。

1. 实体边界检测

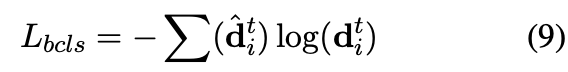
在实体边界检测阶段，对于一个由句子中的第i到第j个词组成的实体，用R(i,j)表示，将ti和tj标注为B和E，区间内的词标注为I，非实体的词标注为O：



对于每个词，通过下面的式子计算和预测其对应的边界标签：

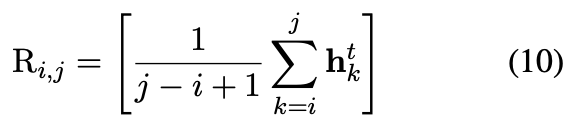


边界检测的损失为：



1. 实体标签预测

得到实体边界信息R(i,j)之后，对于给定输入句子序列X和对于的边界标签序列L，将标签B的每个标签与标签E的标记匹配构造实体区域。考虑到单个词的实体，将标注为B的词语其自身匹配，计算R(i,j)的表示：



对R(i,j)进行实体类别分类：



损失函数为：



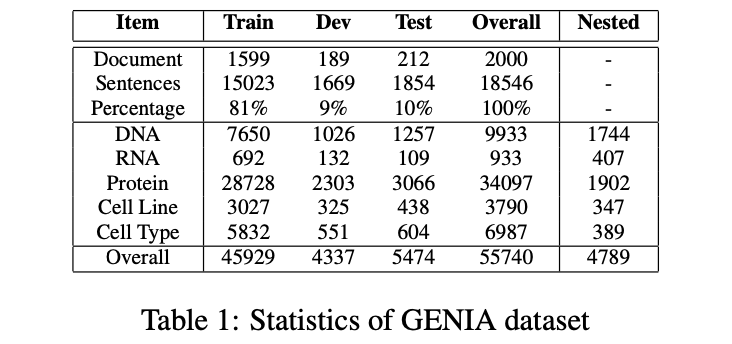
1. 多任务训练

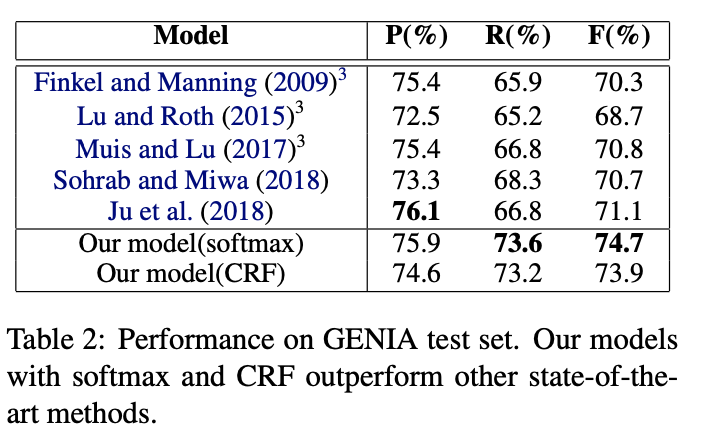
考虑边界检测和实体分类标签预测模块共享同样的实体边界，故将实体检测和标签预测的损失应用于同时训练两个任务，多任务训练的损失函数为：



1. 实验结果及分析

数据集：GENIA





模型解码阶段用了softmax和CRF。对于本实验，softmax跟的表现比CRF稍微好一些，推理的速度也更快。