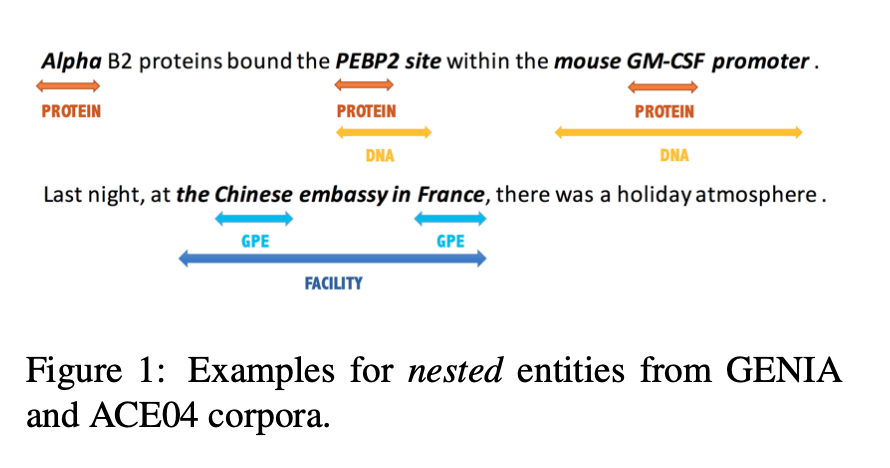
**论文阅读笔记**

A Unified MRC Framework for Named Entity Recognition

1. 贡献
2. 提出一种能够在NER任务中同时处理单位实体和嵌套实体的框架
3. 将NER任务看作一个机器阅读理解任务，而不是序列标注问题
4. 该方法认为提取具有不同类别的两个嵌套实体需要回答两个独立的问题

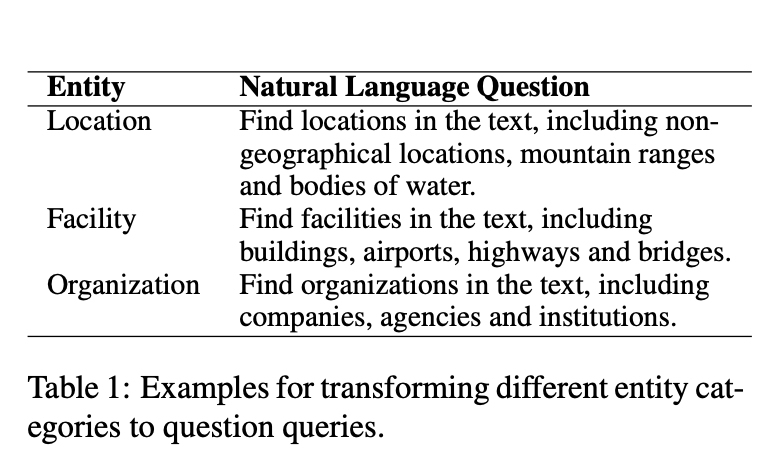


1. 方法
2. 数据集：给定输入序列X={x1, x2, …,xn}，标注序列为Y={y1, y2, …,yn}

将带标注的数据转换为一系列(QUESTION, ANSWER, CONTEXT)形式的三元组，对于y∈Y，它都与一个问题qy={q1, q2, …,qm}有关，其中m表示问题长度。带标注的实体xstart,end={xstart , xstart+1,…, xend-1, xend}是输入序列X的子串，其中start≤end。每个实体都有一个最佳标签y∈Y。基于标签y产生的问题，我们可以得到形如(qy, xstart,end, X)的三元组，分别表示问题、答案和上下文。

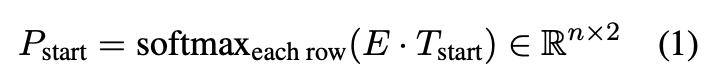
1. 问题的生成

使用标记准则注释作为构造问题的参考，该注释是数据集生成器提供给数据集标注器的准则，他们是类别的描述，精确和通用的表达使得数据很少遇到歧义。

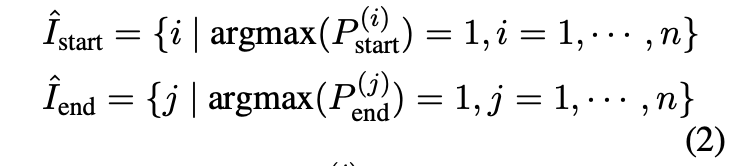


1. 模型细节
2. 给定问题qy，我们要在MRC框架下根据标签y，从X中提取提取文本的范围xstart,end。为了与BERT一致，本文将问题qy与输入X进行拼接，得到字符串BERT将该字符串作为输入，得到上下文表示矩阵E∈Rn×d，其中d表示向量维度。最后要删掉BERT的query向量表示
3. xstart,end范围的选择：完成两个二分类任务，一个预测开始索引，一个预测结束索引

开始索引的预测：给定的向量矩阵E，模型首先预测 每一个词为开始索引的概率：

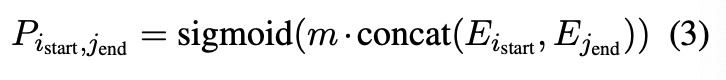
其中Tstart∈Rd×2，是待学习的参数，Pstart的每行表示每个索引为实体起始位置的概率。结束索引也采用同样的方法进行计算。

开始-结束位置的匹配：因为嵌套实体的存在导致在输入X中可能会有多个开始和结束索引被预测出来。故采用将argmax过程应用于Pstar和Pend矩阵的每一行，获得开始位置和结束位置的索引：



其中上标(i)表示矩阵第i行。

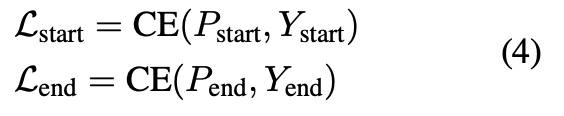
给定开始索引istart和结束索引istart，二分类模型就是预测这两个索引匹配为一个实体的概率：



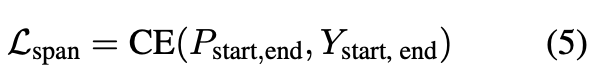
其中m∈R1×2d是待学习的参数。

1. 训练和测试

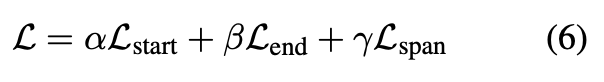
对每个token作为开始项和结束项的损失函数为：



令Ystart,end表示每个开始索引是否与结束索引匹配的最佳标签，起始段索引匹配的损失函数如下：



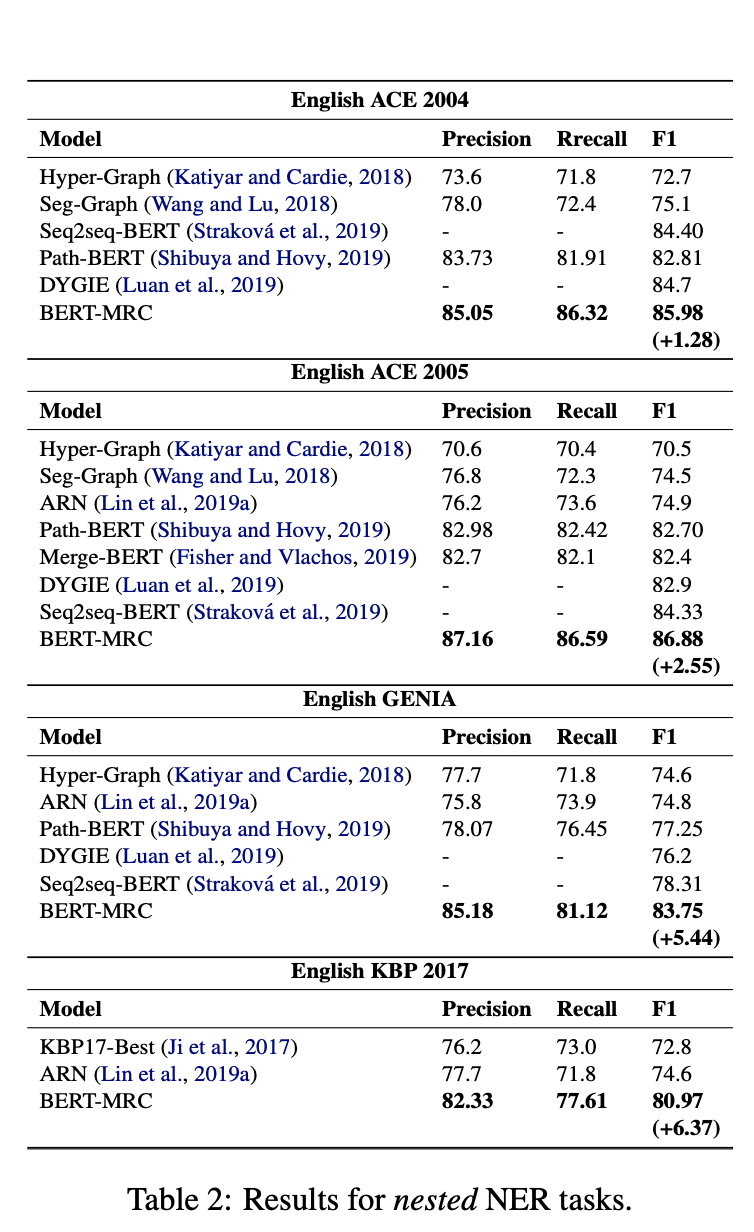
最终的损失函数如下：



其中α，β，γ∈[0,1]，均为控制各个函数对整体训练目标的超参数。

这三个损失函数以端到端的方式实现，均共享BERT层的参数。测试时，首先根据Istart和Iend选择开始索引和结束索引，然后索引匹配模型将提取的起始索引和结束索引对其，从而提取最终答案。

1. 实验结果
2. 嵌套实体NER任务
3. 数据集：ACE 2004, ACE 2005, GENIA, KBP 2017，这些数据集分别包含24%、22%、10%和19%的嵌套实体。
4. Baseline：Hyper-Graph、Seg-Graph、ARN、KBP17-Best、基于BERT的改进模型、DYGIE
5. 实验结果



1. 单位实体NER任务
2. 数据集：English: CoNLL2003、OntoNotes 5.0; Chinese : Chinese MSRA、Chinese OnToNotes4.0
3. Baseline：English：BILSTM-CRF、ELMo、CVT、BERT-Tagger；Chinese：Lattice-LSTM、BERT-Tagger、Glyce-BERT
4. 实验结果

