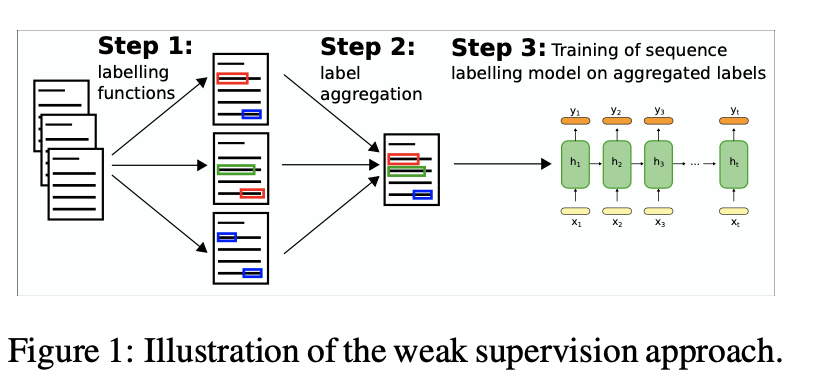
论文阅读笔记一

Named Entity Recognition without Labelled Data: A Weak Supervision Approach

1. 贡献
2. 提出了一个在各种文本域、地名词典、启发式函数和文档级约束下训练的标记功能广泛的神经模型；
3. 构建了一个适用于序列标注任务的弱监督模型，它能够进行概率标记预测；
4. 将标记函数和聚合模型整合为一个开源工具，可将工作扩展到大型数据集。
5. 方法



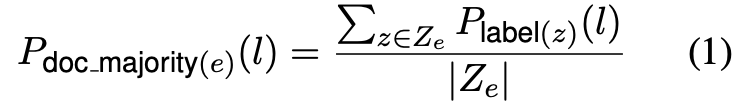
该模型从多个标记函数中收集弱监督信息，每个标记函数将文本文档作为输入，输出一系列与NER标签关联的标记区间。然后通过HMM模型以无监督的方式来汇总这些输出，最后，用这些标签来训练序列标注模型。

标记函数通常用于检测可能标签的子集，产生的结果是概率值而不是确定的值。

**组合标签函数的汇总模型。**

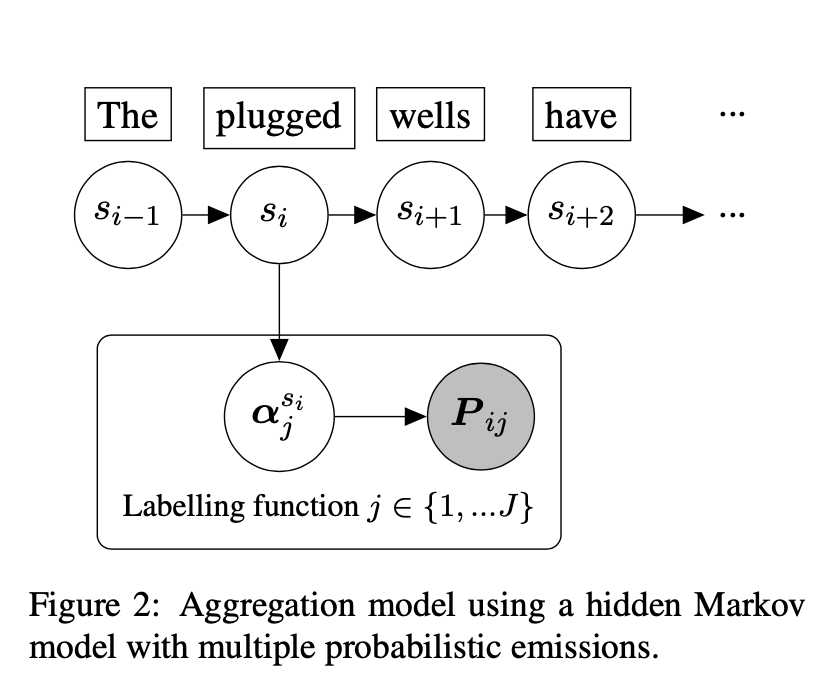
**标签模型**

1. **标记函数**
2. 域外NER模型：通过在有标签数据中训练的序列标注模型所得到的一系列标记函数，所用到的四个序列标注模型都是基于转移的NER模型，该模型通过四个卷积层的堆栈提取特征 ，滤波器大小为3，并包含残差连接，该模型使用注意力特征和一个多层感知机来选择下一个转移。模型用Glove嵌入进行初始化，但该模型不对模型结构和替代方法施加任何约束。
3. 地名词典模型：使用来自维基百科和DBPedia的数据，数据包括国家、语言、民族、宗教或是政治团体。将每个知识库转换为一个Trie数据结构，是用前缀搜索来提取匹配项。
4. 启发式函数：每个启发式函数都用于识别特定类型的命名实体。有几个函数专门用于基于大小写、词性标签或依赖关系的专有名称识别。另外也整合了一系列基于正则表达式的函数，用于检测各种实体的出现，同时也设计了一个专门用于识别日期、时间、金额、百分比和基数/有序值。
5. 文档级别的关系抽取：前面三类标记函数都依赖于数据的局部信息，但是文本具有很强的内部逻辑关联，所以作者引入了一个标记函数去获取文档中标签持续性约束。给定字符串e，在文档中查找所有包含e的字符串，然后标记函数输出文档中该字符串的每个标签的相对频率：

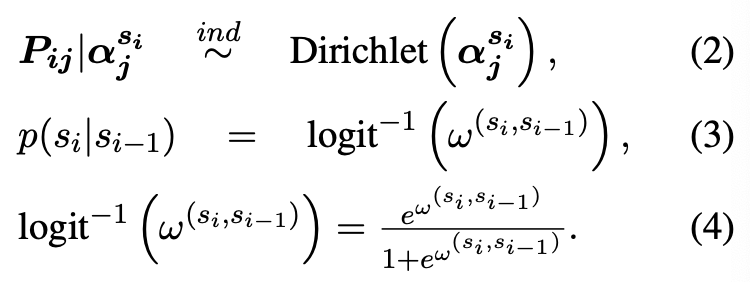


1. **聚合模型**

上述的标记函数的输出随后会通过聚合模型被整合到一个单层注释中。模型以无监督的方式进行工作。



**模型参数：**假设标记函数由J个，NER标签有S个 。聚合模型为HMM，该模型具有多个发射值（每个标记函数一个），假定各个标签互相独立。对每个token的标记i∈{1,…,n}和标记函数j，假设概率标签Pij为狄利克雷分布；对每个潜在状态，si∈{1,…,S}，狄利克雷的参数是独立的向量

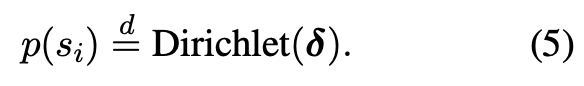


其中w(si,si-1)表示从si-1状态到si状态的转移概率。

**参数评估**：转移概率和与每个标记函数相关的α向量，采用最大期望算法进行参数评估，用前向后向算法计算每一步的期望。

为保障收敛速度，对似然函数引入一个约束，也就是说对每个token i对于的潜在标签si在至少一个标注函数中必须具有非零概率，否则该标签似然性为0，该约束在每个时间步都将状态空间限制为几个可能的标签。

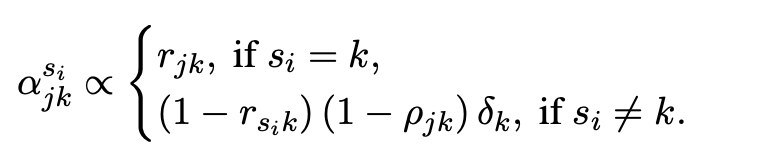
**初始状态分布：**对于最可信的标记函数，潜在状态的初始分布可以定义为基于计数δ的狄利克雷分布：



转移概率也被定义为一个基于频率的狄利克雷函数：



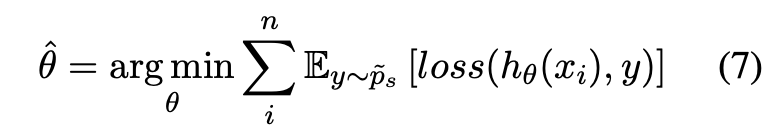
最后为了促进EM算法的收敛，为每个标签函数的发射模型指定信息初始值。假定我们可以对标签k上的标签函数j的召回率rjk和精确度ρjk进行粗略估计，则发射模型参数初始值如下：



解码：参数确定后用前向后向算法将每个token与后验概率关联到可能的NER标签中。

1. **序列标注模型**

将标记函数进行聚合之后，就可以进行序列标注模型的训练，损失函数：

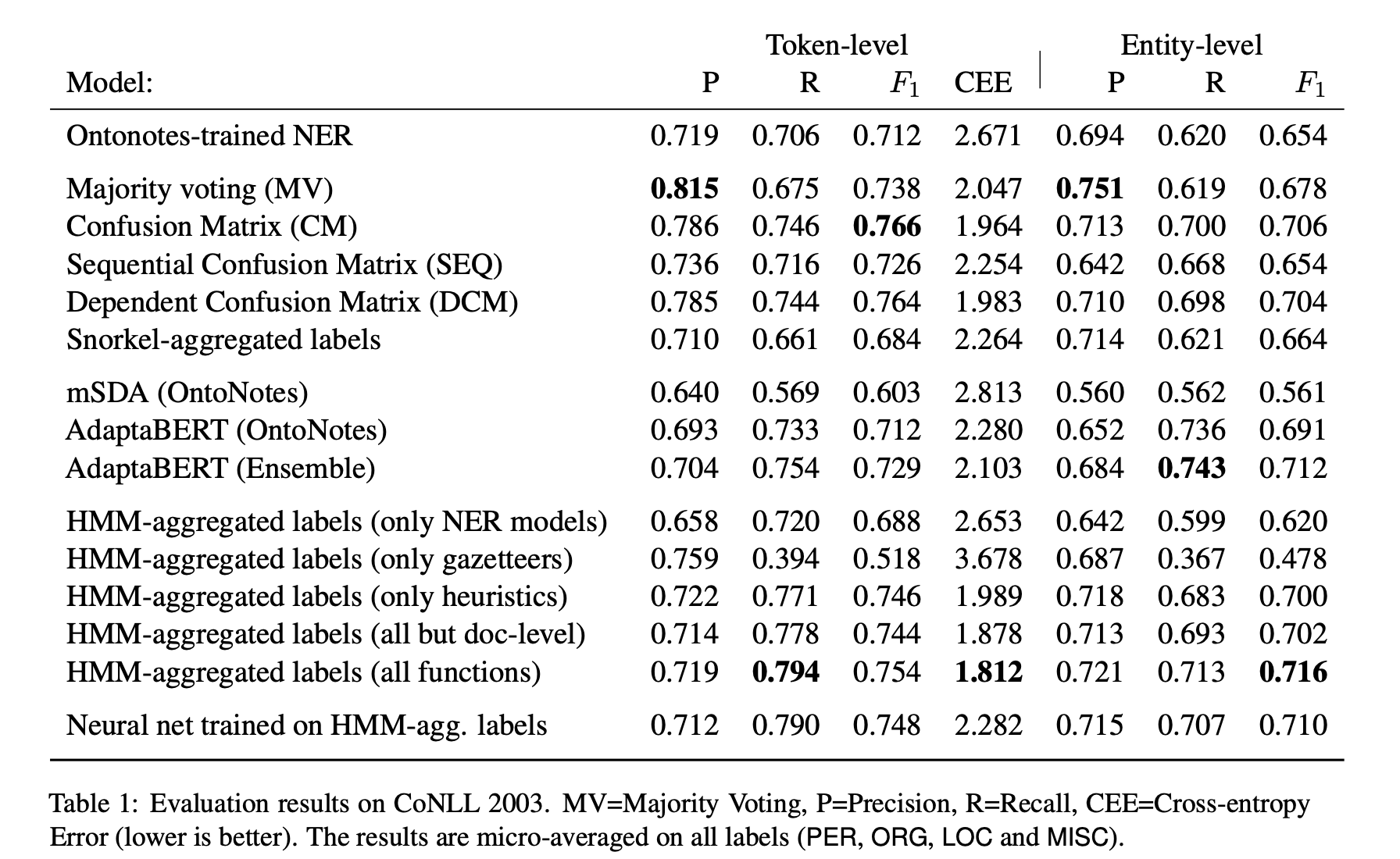


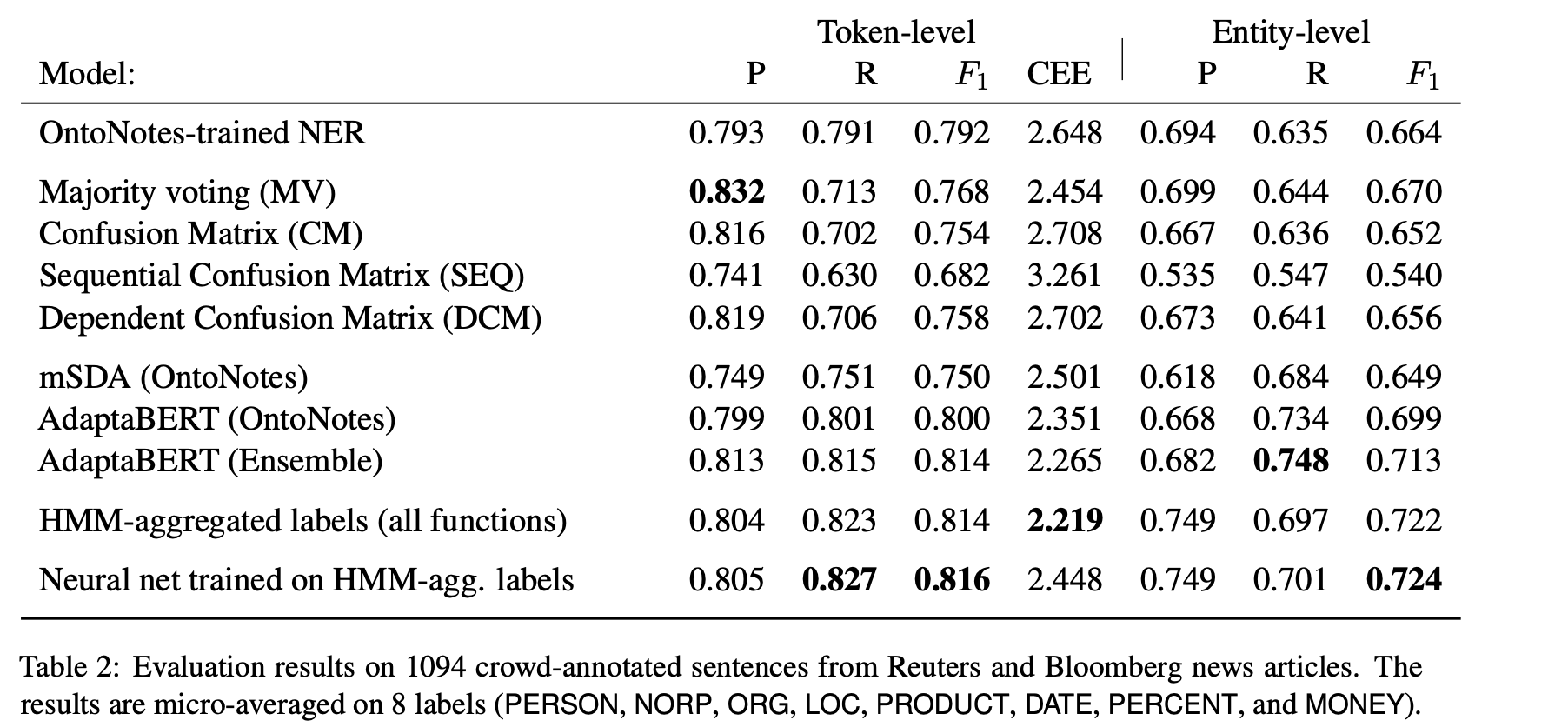
其中hθ(·)是序列标注模型的输出，函数等价于最小化神经模型的输出与聚合模型生成的概率标签最近的交叉熵误差。

1. 实验结果和分析

数据集：CoNLL2003，1163篇文档，35089个实体，包含ORG、PER、LOC和MISC四类实体；Reusters & Bloomberg，1054个句子，包括PERSON、NORP、ORG、LOC、PRODUCT、DATETIME、PERCENT、MONEY、QUANTITY九种实体。

实验结果：





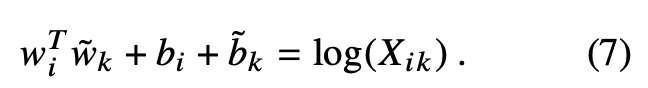
论文阅读笔记二

GloVe: Global Vectors for Word Representation

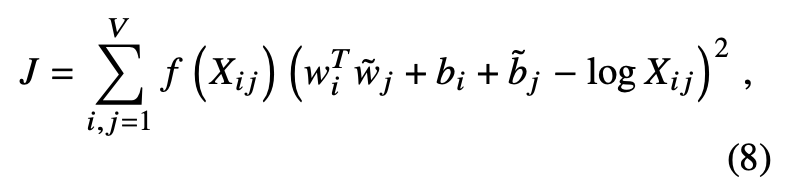
1. 贡献

一个基于全局词频统计的词表征工具，是一个将词表示为含有语义信息的向量的模型。

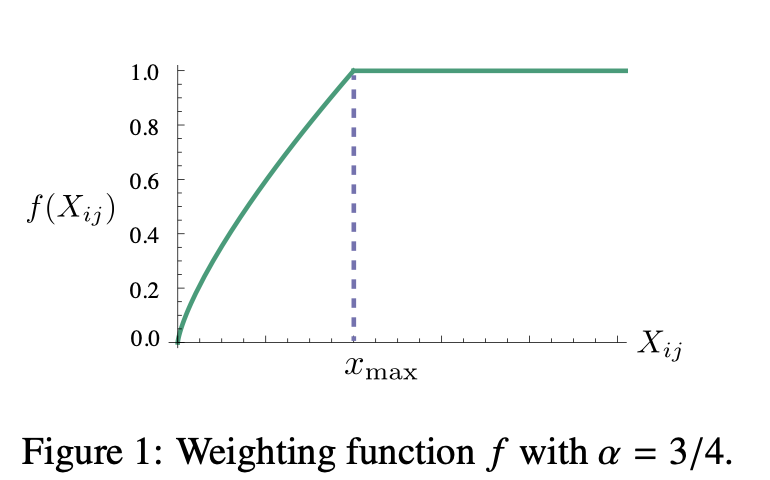
1. 方法
2. 根据语料库构建一个共现矩阵（Co-occurrence Matrix）X，其中元素表示词j与词i在特定大小的上下文窗口(context window)的次数。一般次数为1，Glove根据两个单词在上下文窗口的距离d提出了一个衰减函数：decay=1/d用于计算权重，距离越远的单词所占的总计数(total count)的权重越小。
3. 构建词向量(word vector)和共现矩阵(co-occurrence)之间的近似关系



1. 构造损失函数：



在均方损失函数基础上增加了一个权重函数f(Xij)，用于处理一个语料库中经常一起出现的单词（frequent occurrence），这些单词的权重是有上限的，该函数是非递减的，若两单词没有同时出现，则Xij=0,f(0)=0



作者采用了一个分段函数来描述这种单词的权重，实验中α的取值都是0.75，xmax的取值都是1000。

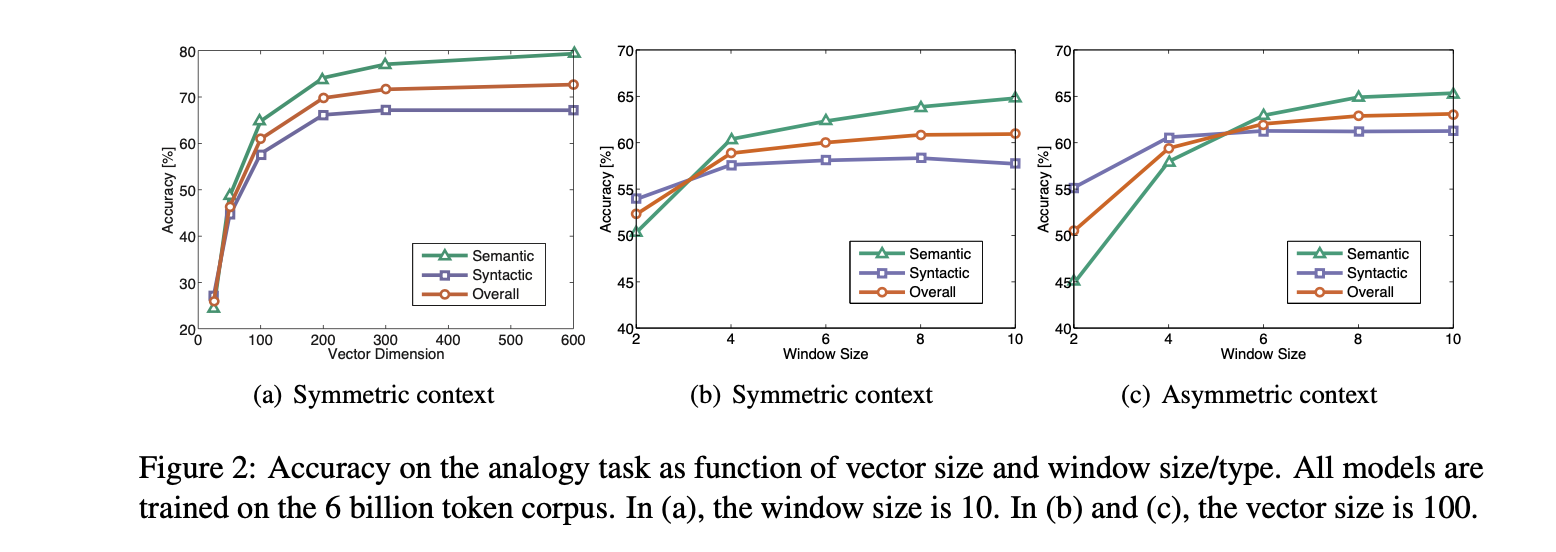
采用AdaGrad的梯度下降算法，对矩阵X中的所有非零元素进行随机采样，学习率设为0.05，在vector size小于300的情况下迭代50词，大于300时迭代100次，直到收敛。

1. 实验结果

训练结果 ：得到两个对称的词向量w\_i和w ̅\_j，最终选择w\_i+w ̅\_j作为最终词向量。初始化值不同相当于加了不同的随机噪声提高结果鲁棒性。

衡量指标：语义准确度、语法准确度以及总体准确度。

结果分析：向量维度300左右最佳，context Window size大致6-10之间。



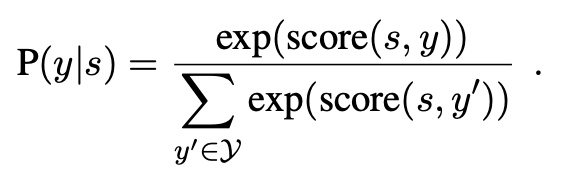
**论文阅读笔记三**

Instance-Based Learning of Span Representations: A Case Study through Named Entity Recognition

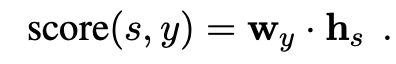
1. 贡献
2. 第一项基于实例的跨度表示学习的工作
3. 基于实例的学习方法能够在不牺牲性能的条件下构建具有高可解释性的模型。
4. 方法
5. 将NER当作一个跨度分类任务

对于给定输入，句子长度为T，首先枚举可能的范围s(X)，然后为每一个范围分配一个标签y。s=(a,b)，1≤a≤b≤T.

每个跨度s被分类为y的概率被定义为一个softmax函数 ：

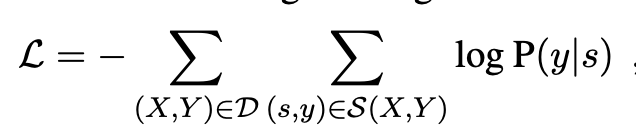


将每个标签的权重向量wy和跨度特征向量hs的内积当作得分：



非实体的标签NULL的得分被设置为一个常量score(s,y=NULL)=0.

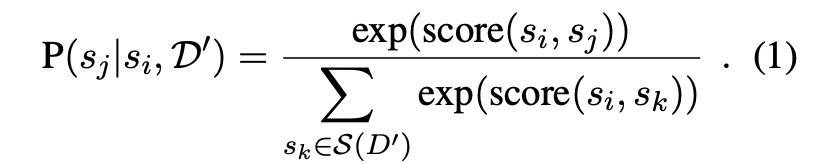
损失函数为负对数似然函数：



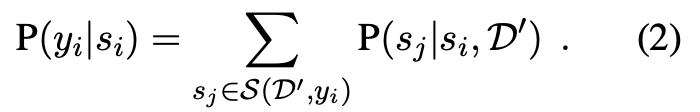
其中S(X,Y)是一系列跨度以及他们的真实标签y。将这种使用标签权重向量的模型用于基于分类器的跨度模型

1. 基于实例的跨度模型

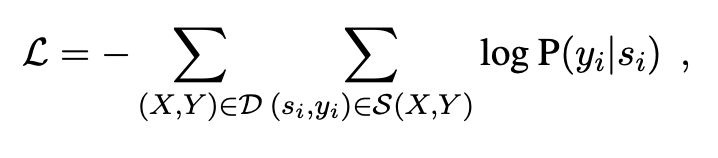
定义邻居跨度概率，即训练集中每个跨度si都会选择一个sj作为它的邻居：



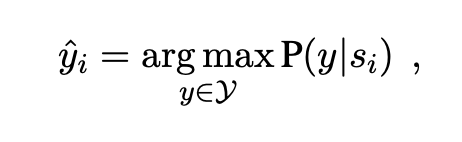
其中，将所有其他的跨度都当作候选，，打分函数返回两个跨度si和sj之间的相似度。然后我们计算跨度si被标记为yi实体的概率：

**

损失函数为负对数似然函数：

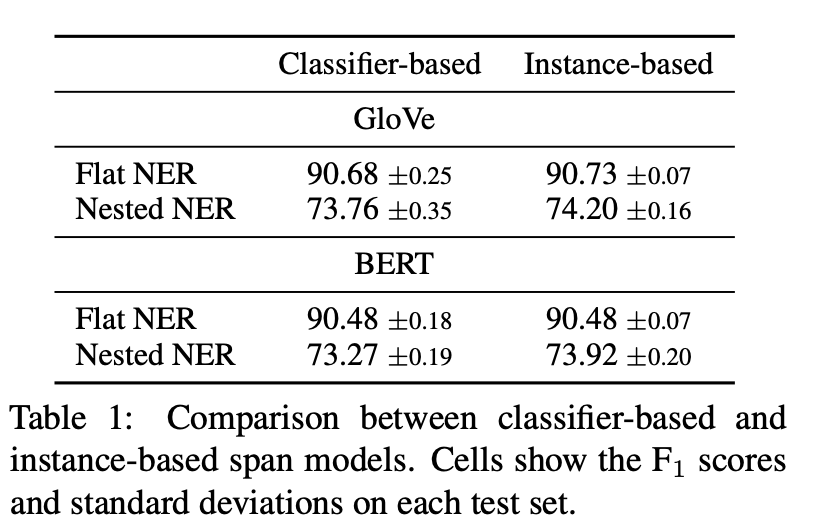
**

最后通过最大化边缘概率得到预测值：

**

1. 实验结果

数据：CoNLL2003和GENIA



1. 思考

在划分单词范围的时候，直接采用了枚举的方式，给定句子长度为T，则计算量为2^T-1，这不是一个非常理智的做法。我感觉可以考虑通过字符级特征和单词级特征来减少待计算的范围的数量。