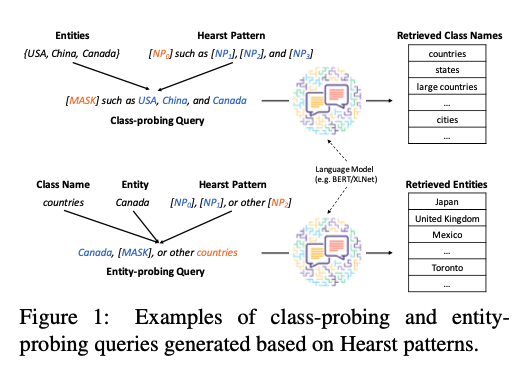
论文阅读笔记一

Empower Entity Set Expansion via Language Model Probing

1. 贡献
2. 本文提出一个新的实体集扩展框架，利用类名来指导整个扩展过程，通过在每次迭代中过滤实体集来解决语义漂移问题；
3. 设计了一个自动生成类名的算法，该算法通过动态监测语言模型来输出高质量的类名
4. 方法



1. 类名生成模块

本模块利用六个Hearst模板去构建类探测查询向量。

Unigram class name: 分别在当前实体集汇总选择三个实体、在六个Hearst模板中随机选择一个模板去构建一个查询向量，比如，选择{“China”, “India”, “Japan”}以及“NPy such as NPa, NPb, NPc”去构建查询模板，“[MASK] such as China, Japan, India”。通过重复这个过程来构建一些列的查询模板，将这些模板送到预训练语言模型中进行训练，获取可能的类名令牌。

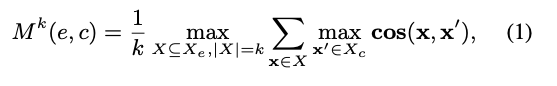
Multi-gram class name: 作者设计了一个光束搜索算法迭代查询预训练模型。首次查询语言模型LM并检索K个最有可能的MASK的单词后，通过在MASK这个token后面分别添加这K个单词来构造新的查询。比如“[MASK] such as China, Japan, India”中，MASK位置可能是countries、nations，将这些可能的单词分别添加到[MASK]后，得到“[MASK] **countries** such as China, Japan, India”和“[MASK] **nations** such as China, Japan, India”，用新的查询去检索语言模型LM，得到单词可能是“Asian”或“Large”，得到更精细的类名，如“Asian Countries”或“Large Countries”。

最多搜索Multi-gram class name三次，并保证所有生成的类名都是名词短语，过滤到无意义字符组合。  
对每个Hearst模板和随机选择到的三个实体，可以得到一系列候选类名。将这些实体集的组合当作候选类名池，用C表示，本模块关注recall而非precision。

1. 类名排序

对上面所生成的类名进行排序，同时加入负样本来过滤错误的实体。

首先引入一个基于语料相似性的措施来计算实体e和类名C之间的相似性。给定类名c，根据六个Hearst模板遮住上义词构建6个实体-探测查询。比如对类名countries，构建“countries such as [MASK]”。查询预训练语言模型LM来获取一系列关于这六个[MASK]的token的词嵌入，该词嵌入表示为Xc。同时用Xe来表示给定语境下实体e的词嵌入。定义实体e和类名c之间的相似度为：



其中x表示实体e的向量，x’表示类名c，cos(x, x’)表示两个向量的余弦距离，内层max取实体e和实体-探测查询之间相似度最高，外层max操作取前k个与e最相似的查询，然后取他们的平均值作为实体e和类名c之间最终的相似度。

对实体集E中的每个实体，每个实体都有一个类的排序列表，L1, L2,…,L|E|。最后根据每个实体的类得分列表Li来整合当前实体集的类得分，选择排名第一的类为正类，表示为cp。



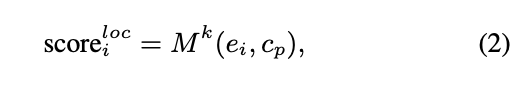
为约束语义，同样为目标类选择一组负类名，假设初始种子实体集为E0，选择在所有跟种子实体有关的、得分低于cp的类名，令其为{Li|ei∈E0}，将他们作为负类名，表示为CN。

1. 利用类名进行实体扩展迭代

利用前两个模块得到的正类和负类进行实体集扩展。提出一个新的排名组合算法进行实体选择。

首先引入两个打分函数，

第一个打分函数利用正类计算每个实体的得分：

其中M来自等式(1)表示实体相似性，因为这个得分仅考虑了与ei相似的前k个最佳分类结果，local表示局部得分。

第二个打分函数根据上下文表示计算当前实体集中每个候选实体和现有实体之间的相似度：

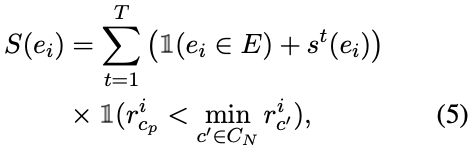
其中vei表示候选实体集中实体的上下文表示，ve表示现有实体的上下文表示。Es表示从实体集中采样的少数实体，一般为3个。随机采样的目的是为了缓解由于负例的存在导致的语义漂移问题，因为该函数利用了词的上下文无关表示，更好地反映了所有实体之间的相似度，故该分数为全局得分。

用前面两个函数的几何平均表示候选实体的得分：



**新的排名组合算法：**

首先重复从当前实体集E中采样T次计算等式(3)的全局得分，得到T个实体排名列表；然后根据类名排序过程去获取实体集E的类排序列表，属于目标语义类的实体应满足(1)出现在多个实体排名列表中的顶部位置以及在类排名列表中最佳类cP排名应高于负类cN：

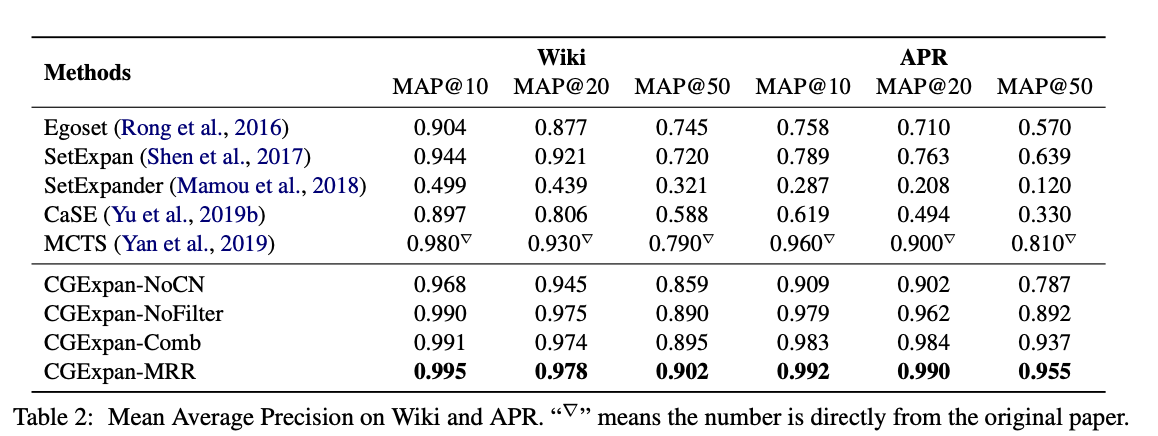


其中，1(·) 是个指标函数，r是类名c在实体集ei中的排名列表。

整个模型从种子实体开始，迭代运用前面三个模块获取实体排序列表，直到达到下面两个条件则停止：(1)已扩展的实体集达到预设的个数;(1) 实体集不再扩展

1. 实验结果

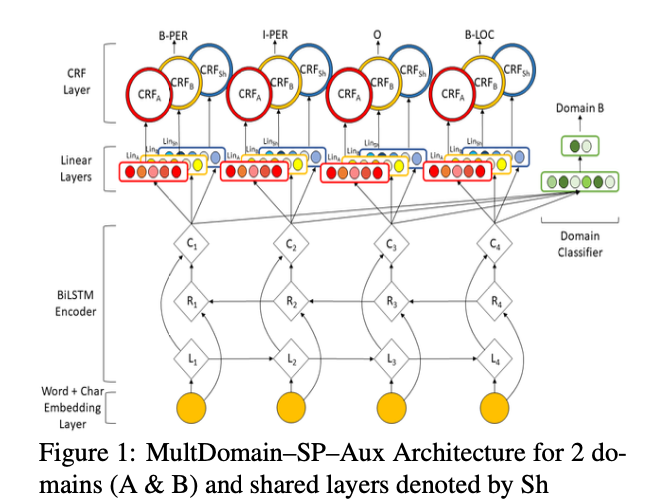
数据集：APR，Wiki



**论文阅读笔记二**

Multi-Domain Named Entity Recognition with Genre-Aware and Agnostic Inference

1. 贡献
2. 提出了一种基于BiLSTM-CRF框架的命名实体识别神经网络
3. 引入了多领域、无标签多领域和未知流派三种实验配置进行训练
4. 捕获多个领域数据的公有特征，也记录特定领域的私有特征
5. 方法

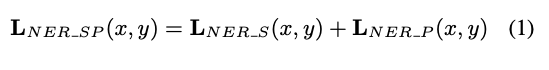


整体结构如图一所示，输入为数据的词向量和字符向量，经过BiLSTM编码获取上下文特征，BiLSTM编码结果送到一个线性前馈网络层中，将结果映射到标签分布中，最后用CRF对这些标签进行建模。CRF考虑标签间的依赖性，也对输出所遵循的约束进行建模。

**Proposed Architecture (MultDomain–SP–Aux)**

1. 处理BiLSTM输出的特定领域和独立的特征的前馈层；
2. 特定域和独立信息的前馈层+CRFs；
3. 标签建模.

共享参数层损失函数和私有参数层损失函数：



域标签的多任务学习：



1. 实验

数据集：CoNLL2003、Twitter、ontonotes三个领域内有标记的数据集，Zero Shot Genres无标注非特定领域数据集，包括新闻、字幕和其他文档数据，用BIO标注法将四个数据集的都被标记为同样的实体类型，使用相同的实体类型和相似的准则进行标注。

为了在四个数据集上都得到可比较的结果，只预测ORG(GPE\facilities)、PER和LOC三种类型的实体，忽略其他内容。

