**周报**

## 1.本周任务

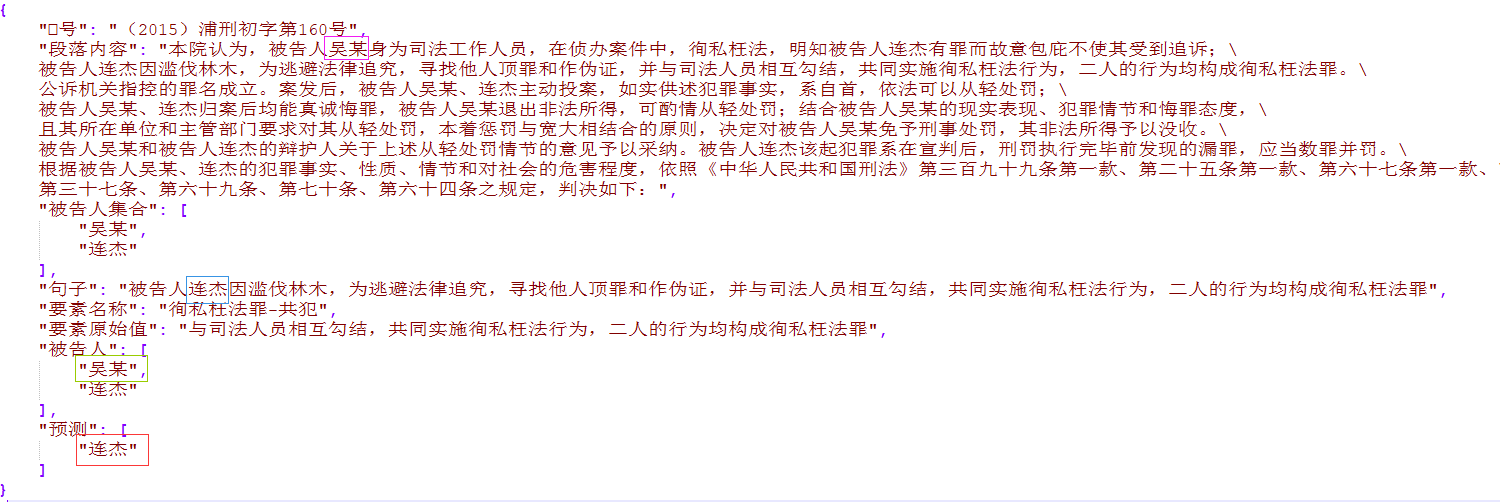
（1）论文修改

本周主要完成对实验部分的改进，对于之前的实验模型结构进行微调，在embedding层换用几种不同的词语向量表示模型，在最终的链接环节尝试使用GCN进行linking，这周主要在进行代码层面的修改，对论文的模型进行了微调。

（2）法律竞赛

一、当前进度

根据上周的讨论结果，数据判错的原因大部分在于句子中信息不足，有许多被告人无法仅仅根据句子中的信息判断归属。如下图数据所示，要素名称为“徇私枉法共犯”，说明应该是两个人徇私枉法，但是句子中仅仅包括被告人连杰的信息，如果要找到关于吴某的徇私枉法的信息，需要去段落中才能发现吴某和连杰是互相徇私枉法。



所以主要改善方向是去段落中寻找信息，对句子的信息进行补全，以下是信息补全策略尝试（皆使用原来的Classify+ERNIE模型 ）：

测试样例数6959

句子无补充内容：句子缺失被告人样本数808，本地F1=0.972，提交F1=0.891

|  |  |  |
| --- | --- | --- |
| 句子补充内容 | 句子缺失被告人样本数 | 提交F1 |
| 段落前200字符 | 271 | 0.84 |
| 段落前100字符 | 361 | 0.86 |
| 段落前50字符 | 440 | 0.893 |
| 使用段落前字符补充，直到509字符 | 216 | 0.84 |
| 使用段落中句子位置前10字符和后20字符 | 703 | 0.893 |
| 使用段落中句子位置前5字符和后15字符 | 707 | 0.892 |
| 当被告人不在句子中时，找到段落中该被告人所在的子句，长度为50 | 0 | 未提交，本地运行效果不如无补充。 |

根据测试结果，大多数信息都在段落的前200个字符以内，但是如果单纯的将信息都补全到句子上，那么结果会因为添加了很多噪声的原因不增反降，而段落前50个字符就可以涵盖存在被告人不在句子中的一般的样本，效果要比单纯的使用句子要稍微好一些。所以目前的主要方向还是怎么让数据集的句子信息能够在保持字符数比较少的情况下涵盖尽可能多的被告人信息。

二、改进方向

1、继续在数据集上进行句子信息补全，使用以下几种策略：

（1）当被告人不再句子中时，找到段落中该被告人的位置，去除该被告人临近的所有被告人，之后截取一定数量的子句进行拼接。

（2）对（1）修改，因为段落中找到的子句和给出的要素原始值不一定相同，所以使用要素名称可能会有更好的效果。

（3）对（1）修改，因为去段落中寻找信息找到的更多的是噪声，所以对找到的信息进行剪枝。找到被告人的位置之后不是对句子进行拼接，而是和要素原始值进行最大子串比较，找到和要素原始值匹配性最高的一个句子和原始句子进行拼接或者单纯使用找到的句子。

2、除了对句子信息进行补全，也可以在预测的时候进行改进，比如：

（1）和规则进行结合，如果当前被告人不在句子中，则使用规则进行判断。

（2）同样是和规则进行结合，规则的主要问题在于规则的不完全正确性和要素原始值不在原句子中，所以对于要素原始值不在句子中的情况下，可以使用模型进行预测。

（3）对于不在句子中的被告人，使用滑动窗口法对原段落内容进行滑动比较，只要有一个span的子句和被告人匹配，那么将其作为正确被告人，如果所有的span都不匹配，那么将该被告人判定为错。

（3）论文阅读

《Zero-Shot Entity Linking by Reading Entity Descriptions》

Abstract：

提出了zero-shot实体链接任务，其中文本中提到的内容必须链接到不带域内标记数据的不可见实体。目标是实现到高度特殊化的域的迁移，因此不假设元数据或别名表。在此设置中，实体仅由文本描述标识，模型必须严格依赖语言理解来解析新实体。首先，证明了在大型的未标记数据上预先训练的强阅读理解模型可以被用来推广到不可见的实体。其次，提出了一种简单有效的自适应预训练策略，称之为域自适应预训练（DAP），以解决新域中与未知实体链接相关的域迁移问题。在为这项任务构建的新数据集上进行了实验，结果表明，DAP比包括BERT在内的强预训练基线有了改进。

Contributions

1.我们提出了一个新的zero-shot实体连接任务，旨在挑战最小假设下实体连接系统的泛化能力。我们为这个任务构建了一个数据集，后续将被公开。

2.我们使用最先进的阅读理解模型建立了一个强大的基线。我们表明，在上下文中的mention和实体描述之间的attention对于这项任务是至关重要的，这在以前的实体链接工作中没有使用过。

3.我们提出了一种简单而新颖的自适应策略，称为域自适应预训练（DAP），并证明它能进一步提高实体链接性能。

任务描述：实体链接（Entity Linking）就是在某些数据库中把文本中对应的实体找出来，通常，这个并不直接出现，而是以代词或其他形式出现。比如，现在给定一个指定m（mention）和它的上下文，我们想要在给定的实体集合中找出所有代表的实体。

实际上，许多实体链接系统依赖于以下资源或假设：

l 单个实体集

这假设在训练和测试示例之间共享一组完整的实体。

l 别名表

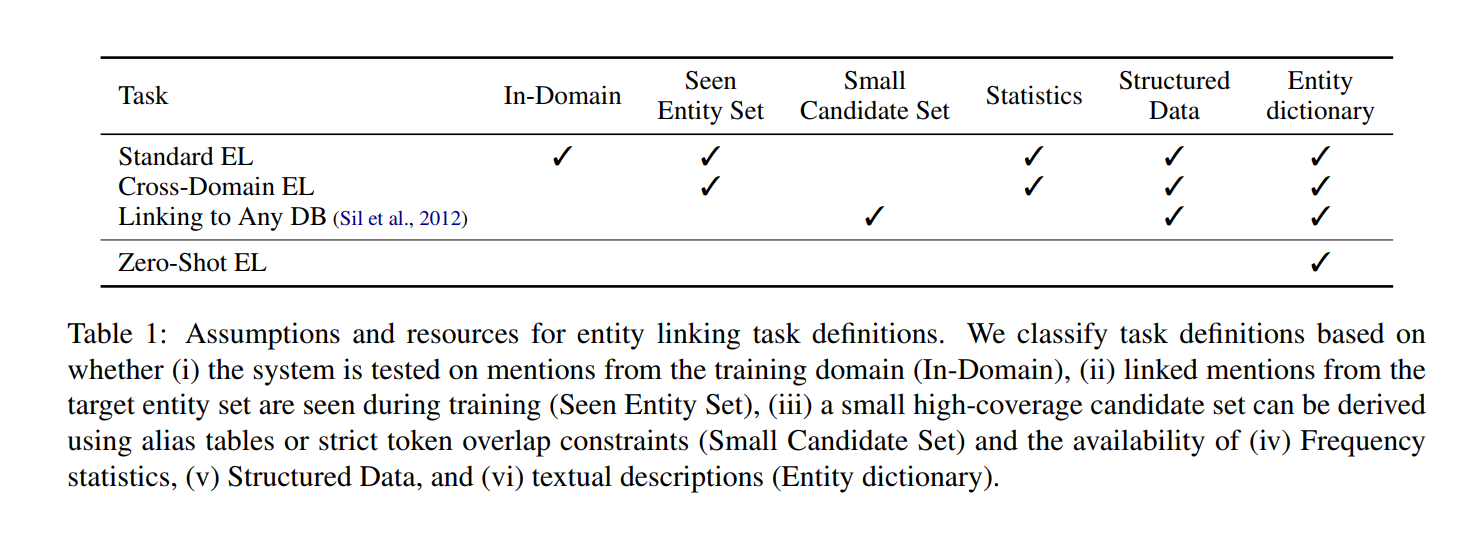
别名表包含给定提及字符串的候选实体，并将可能性限制为相对较小的集合。这些表通常是从标记的训练集和特定于域的启发式方法编译而来的。

l 频率统计

许多系统使用从大型标记语料库中获得的频率统计来估计实体的受欢迎程度和提到字符串链接到实体的概率。这些统计数据在可用时非常强大。

l 结构化数据

有些系统假设可以访问结构化数据，例如关系元组（例如，Barack Obama、souther、Michelle Obama）或类型层次结构，以帮助消除歧义。



本任务的假设：

实体字典E= {（EI，DI）}， i＝1，..，k的存在，其中DI是实体EI的文本描述。

本任务的目标是建立实体链接系统，可以推广到新的领域和实体词汇，称之为worlds。

定义：M:mentions；U:documents；E:entities



Dataset Construction：

构建了一个新的数据集来研究使用Wikia文档的zero-shot实体链接问题，基于超链接自动提取标记的mention。mention和实体具有丰富的文档上下文，可以通过阅读理解方法加以利用。每个Wikia都有大量与特定主题相关的唯一实体，之后用于评估实体链接系统的领域泛化能力。

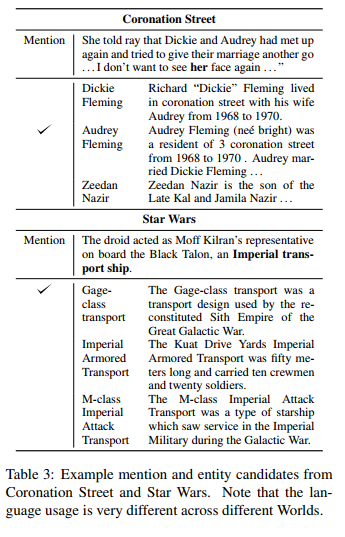
本文使用了Wikia来构建16个不同类别的数据集，8个用于训练，4个用于验证，4个用于测试。对应抽取的指定（mention），它与它对应的实体有三种可能：

l 高覆盖（High Overlap）：mention和对应实体完全一致，约占5%，这部分会降低采样。

l 多类别（Multiple Categories）：实体在mention后再跟一个短语，约占28%

l 歧义子串（Ambiguous substring）：mention是实体的一个子串，约占8%

l 低覆盖（Low Overlap）：其他情况，占59%



Models for Entity Linking

本文的实体链接分为两个步骤，一个候选集生成，二是候选集排序。

对于候选集生成（Candidate generation），本文使用了BM25方法（TF-IDF的变体）来度量mention和候选文档的相似度。再用Lucene打分并选前K个实体作为候选集。其中在本文中K=64，平均覆盖率略低于77%，说明在候选集产生阶段还有很大的改进空间。

对于候选集排序（Candidate ranking），与Bert相同，本文把mention m和实体描述e连接起来并作为序列作为输入送入模型，即[CLS]m[SEP]e[SEP]，深层Transformer 编码器（12层）将其产生为一个向量hm,e，然后用一个参数w进行打分wThm,e，损失函数使用的是softmax，这个方法为Full-Transformer。此外，还有两个变体：

l Pool-Transformer：这里mention和实体描述被分别送入Transformer，即[CLS]m[SEP]和[CLS]e[SEP]，得到两个向量hm和he ,得分为hTmhe 。

l Cand-Pool-Transformer：同样使用两个Transformer，但是加入了一个新的attention模块，用he来关注m的每一个词语表示。

Adapting to the Target World

希望用预训练的方法提高模型的泛化性，这有三个方法：

l Task-Adaptive：在无标注的训练领域和目标领域上的数据联合预训练，然后用标注的训练领域数据微调。

l Open-Corpus：使用通用的大规模无标注语料训练，如BERT，ELMo；然后用标注的训练领域数据微调。直觉上来讲，如果大规模语料足够大，目标领域应该是该语料的一部分。

l Domain-Adaptive pre-training（DAP）：除了其他方法的预训练阶段外，我们还建议插入倒数第二个域自适应预训练（DAP）阶段--模型只在目标领域数据上预训练，然后在训练领域数据上微调，这是为了优先学习目标领域的知识。

Experiments：

无监督训练：Masked LM

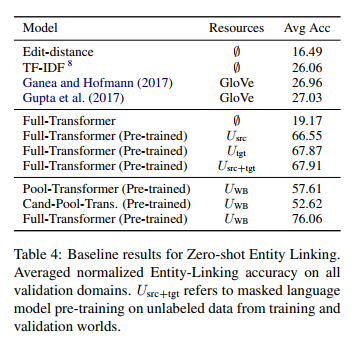
Evaluation

normalized entity-linking performance：测试子集中在候选生成期间检索到的前k个候选中包含gold实体的性能。

unnormalized performance：在整个测试集上进行评估。

我们基于IR的候选集生成在验证集和测试集上的top-64召回率分别为76%和68%。因此，unnormalize性能的上限由这些数字限定。

Result：



说明了pre-training的重要性。Full-Transformer和其他变量之间的显著差距表明了通过Transformer中嵌入的交叉注意机制允许两个输入之间进行细粒度比较的重要性。

## 2.下周任务

（1）继续修改论文

（2）法律竞赛推进

1、继续在数据集上进行句子信息补全，使用以下几种策略：

（1）当被告人不再句子中时，找到段落中该被告人的位置，去除该被告人临近的所有被告人，之后截取一定数量的子句进行拼接。

（2）对（1）修改，因为段落中找到的子句和给出的要素原始值不一定相同，所以使用要素名称可能会有更好的效果。

（3）对（1）修改，因为去段落中寻找信息找到的更多的是噪声，所以对找到的信息进行剪枝。找到被告人的位置之后不是对句子进行拼接，而是和要素原始值进行最大子串比较，找到和要素原始值匹配性最高的一个句子和原始句子进行拼接或者单纯使用找到的句子。

2、除了对句子信息进行补全，也可以在预测的时候进行改进，比如：

（1）和规则进行结合，如果当前被告人不在句子中，则使用规则进行判断。

（2）同样是和规则进行结合，规则的主要问题在于规则的不完全正确性和要素原始值不在原句子中，所以对于要素原始值不在句子中的情况下，可以使用模型进行预测。

（3）对于不在句子中的被告人，使用滑动窗口法对原段落内容进行滑动比较，只要有一个span的子句和被告人匹配，那么将其作为正确被告人，如果所有的span都不匹配，那么将该被告人判定为错。

（3）继续阅读few-shot论文。