**周报**

## 1.本周完成任务

（1）修改视觉实体链接实验

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，由于更换部分模型所以需要一些时间进行代码的调试，争取获得更好的实验效果。

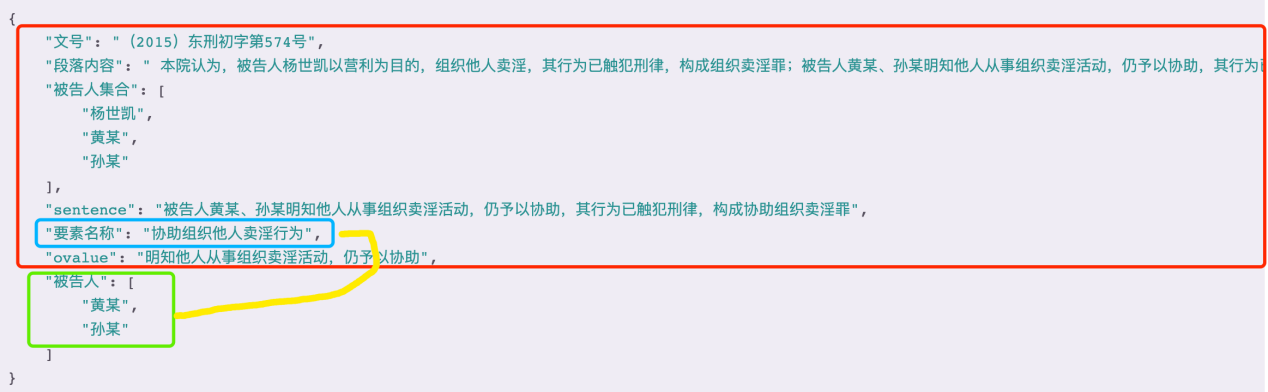
（2）法律竞赛

1. **比赛介绍：**

**任务：**

对输入的文本、要素与当事人进行匹配判断，判断在当前输入文本中，要素与当事人之间的对应关系。

**示例：**



具体来说，就是给定红色框内的输入，里面包含了蓝色框的要素，目的就是输出哪些人参与了这个要素。 比如这里有三个被告人，“杨世凯”是“组织卖淫罪”；“黄某”“孙某”才是“协助组织他人卖淫行为”。所以输出“黄某”“孙某”。

1. **当前进度：**

1、官方给出的baseline使用的是ner+ernie的方法，text\_a是要素原始值，text\_b是句子，模型找出text\_b中所有的与text\_a(ovalue)对应的被告人实体，本地测试集上跑出来f1为95，平台最终结果是88.724。

2、我们当前结果最好的是使用classify+ernie的方法

具体操作：将被告人集合中每个人都拼接到ovalue前面作为text\_a，将句子作为text\_b,然后进行文本相似度判断分类，text\_a与text\_b相似度高的话分类为1，此时被告人加入最终结果集。

结果：测试集上跑出来f1为97，平台最终结果是89.177，排名第二，第一名是89.599。

1. 我们发现数据集中存在着某些规则，如果单纯使用规则就可以跑出84的结果，如果使用规则对模型的预测结果进行后处理，也许会有比较好的结果。
2. **改进想法：**

1、现在模型的P比R要高一些，也就是说需要改进的地方在找到更多的被告人，模型后处理可以在这个方向上进行。

2、接下来打算尝试使用其他模型比如bert、roberta等模型对ernie模型进行替换比较，找出一个效果最好的模型。在此基础上对分类模型进行规则上的后处理，提高泛化性。

3、同时打算使用其他方法与分类方法进行比较，比如阅读理解等。

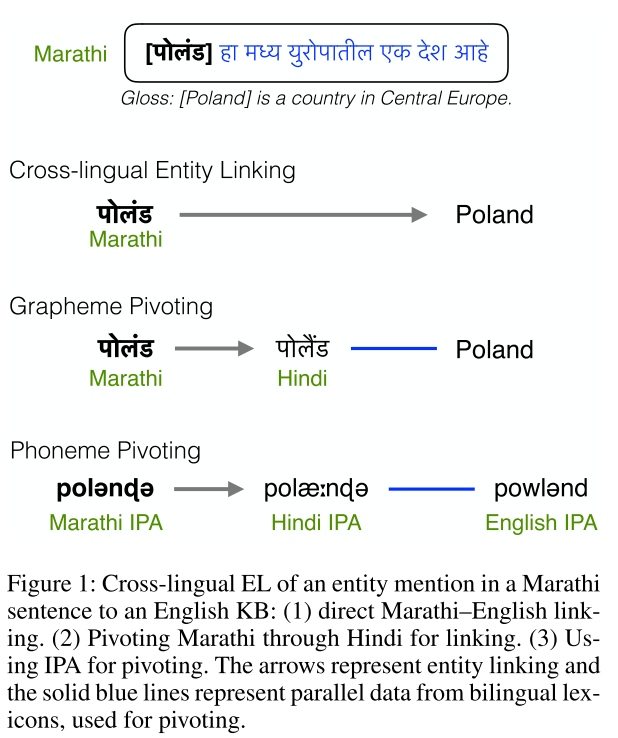
1. 最后尝试使用paddlepaddle的动态图替换静态图。

（3）论文阅读

《Zero-Shot Neural Transfer for Cross-Lingual Entity Linking》

跨语言实体链接

跨语言实体链接将源语言中提到的实体映射到另一种(目标)语言的结构化知识库中的相应条目。以往的研究主要依靠双语词汇资源来弥补源语言和目标语言之间的差距，但是对于许多低资源语言来说，这些资源是稀缺的或不可用的。举个例子，马拉地语(“波兰”)作为输入实体，链接到英语知识库(KB)中的相应条目。



相关工作

在单语言中，字符串相似度和Wikipedia锚文本可以很好的用来进行实体链接。但在跨语言中，这行不通，因为跨语言的实体大部分是不相似的。对于跨语言实体链接任务，(Tsai and Roth 2016; Pan et al. 2017; Tsai and Roth 2018)使用双语资源来弥补这一差距，包括词汇和维基百科的跨语言链接。然而,世界上绝大多数的语言资源缺乏。为了这种低资源语言(LRLs)也可以进行跨语言实体链接，必须设计不过多依赖LRLs中的词汇或其他资源的方法。

自2011年以来，tac-kbp在实体链接方面的任务一直以中文/西班牙语链接到英语为特色(Ji, Grishman, and Dang 2011; Ji, Nothman, and Hachey 2014; Ji et al. 2017)。McNamee et al. (2011) 介绍了跨语言实体链接任务，并设计了一种基于Wikipedia语言链接的候选实体检索技术。Tsai and Roth (2016) 在EL上对12个语言使用词嵌入。Pan et al. (2017)在282对语言对的EL中，使用逐字翻译进行大量的多语言链接。Tsai and Roth (2018)提出更好的名称翻译（name translation），以提高现有基于翻译的EL技术的性能。Sil et al. (2017)提出的神经模型基于多语言词嵌入和维基百科链接，在TAC2015数据集上实现了SOTA的结果。

本文工作

本文提出了pivot-based entity linking(PBEL)：基于低资源语言有紧密相关的高资源语言（HRLs）。比如马拉地语和老挝语资源相对较少，但与资源丰富的北印度语和泰语属于同一语系。本文利用这些HRLs中提供的双语词汇和结构化信息来改进LRLs的实体链接性能。

算法创新

Zero-shot transfer of neural entity linking models：

使用一个高资源语言HRL和英语之间的双语词典，训练一个将HRL中的实体链接到英语知识库中的字符级别的神经模型。这个模型可以迁移到为LRL执行实体链接，不需要有任务具体的语言微调。这种迁移学习方案在密切相关的语言之间使用时，已经成功地应用于其他任务，如形态标记和机器翻译(Zoph et al. 2016;Cotterell and Heigold 2017)。

Pivoting：

实体并不是直接从一个LRL链接到英语，而是先链接到一个密切相关的HRL。然后，使用HRL中现成的双语词汇来获得相应的英语实体链接。即使用HRL作为LRL和英语之间的一个中间轴。

The use of phonological representations for cross-lingual EL：

当HRL和LRL不使用同样的书写系统的时候，字符级别的模型迁移将会失败。使用国际音标表（IPA）来建立不同语言之间的联系，如前面的图中”Poland”在马拉地语和印度语中的发音是非常相似的。

本文提出两种表示：

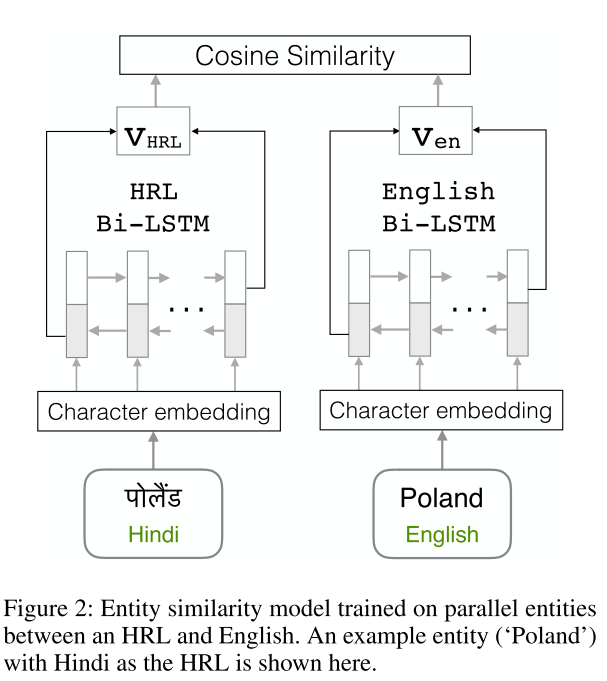
Phoneme embeddings音素嵌入

使用Epitran将HRL和英语之间所有相对应的训练数据转换为IPA, Epitran是一个图形-音素系统，支持超过55种语言(Mortensen, Dalmia, and Littell 2018)。

Articulatory feature embeddings发音特征嵌入

使用PanPhone转化IPA训练数据为发音特征(Mortensen et al. 2016)。发音特征可以潜在地捕获在IPA中可能不明显的发音的重要特征，正如改进低资源命名实体识别(Mortensen et al. 2016)所展示的。

模型框架：



2.下周工作计划

（1）完成视觉实体链接的后续实验，对比之前的模型是否存在改进，如果改进的话对相关部分进行论文修改。

（2）王硕的任务主要是法律竞赛方面的辅助。

1、现在模型的P比R要高一些，也就是说需要改进的地方在找到更多的被告人，模型后处理可以在这个方向上进行。

2、接下来打算尝试使用其他模型比如bert、roberta等模型对ernie模型进行替换比较，找出一个效果最好的模型。在此基础上对分类模型进行规则上的后处理，提高泛化性。

3、同时打算使用其他方法与分类方法进行比较，比如阅读理解等。

4、最后尝试使用paddlepaddle的动态图替换静态图。

（3）阅读相关的zero-shot entity linking论文。