

****

本科毕业设计（论文）

中期检查报告

**基于深度学习的专利与标准的关联关系发现技术**

|  |  |
| --- | --- |
| 姓名： | **霍飞烨** |
| 学号： | **18231098** |
| 学院指导老师： | **张辉** |
| 院系 | **高等理工学院** |

2022年 4 月 8 日

目 录

[1. 课题简介 1](#_Toc98168773)

[1.1 课题背景 1](#_Toc98168774)

[1.2 研究目标和内容 1](#_Toc98168775)

[2. 论文工作进展 1](#_Toc98168776)

[2.1 总体研究进展 1](#_Toc98168777)

[2.2 总体研究方案 1](#_Toc98168778)

[2.3 主要研究成果 2](#_Toc98168779)

[2.3.1 \*\*\*研究成果 2](#_Toc98168780)

[3. 尚存的问题及后期工作安排 2](#_Toc98168781)

[3.1 尚存的问题及措施 2](#_Toc98168782)

[3.2 后期工作安排 2](#_Toc98168783)

[4. 参考文献 2](#_Toc98168784)

# 课题介绍

## 1.1 课题背景

前一段时间，世界知识产权组织（WIPO）公布了 2021 年PCT国际专利申请排名榜单，申请量排名前 5 位的国家分别是中国、美国、日本、韩国和德国，其中中国企业专利数量共计 69540 件，同比增长 0.9%，连续第三年位居申请量排行榜第一位。爆炸式增长的专利背后，是繁复严谨的技术标准的支持与约束。但是随着新的技术领域的不断开拓，旧有的技术标准面临着更新与完善的问题。这就需要我们分析与相关领域的专利相关联的技术标准。

在此背景下，山西省“十四五”规划和山西省委《关于实施“111”创新工程支撑引领高质量转型发展的意见》（晋办发[2020]5号）、山西省政府《实施“111”创新工程支撑引领高质量转型发展工作方案》（晋政办发[2020]27号）文件精神，聚焦“111”创新工程的标准化需求，全面调研国内外标准化及专利现状，研究形成科技创新所需的标准化及专利数据集成方案，集成与产业密切相关的标准化及专利数据信息，产出1-2个优势重点技术领域的标准、专利内容揭示数据集，提出“111”创新工程的标准及专利创制服务体系建设方案，通过构建知识重组能力来加快技术标准及专利形成，以标准数据及专利数据形成的知识库服务促进山西省科技成果产业化，以标准化服务助力山西省打造一流创新生态，为实现山西省经济社会高质量转型发展提供标准化支撑。

传统的标准化分析方法，需要相关的专家人工分析与新增专利的相关联的技术标准，从而再根据提炼的结果来制定相关的决策。在过去专利数量增速未曾如此迅猛的情况下，这种方法未尝不是一种解决问题的有效方式。但是随着信息时代的到来，与专利相关的信息也正式步入大数据时代，此时仍然依赖人工分析数据并制定决策便显得低效且不可避免地导致决策准确率的下降。因此，我们亟需一个高效且准确的方法来对日新月异的技术专利与现有的技术标准进行数据挖掘。

## 1.2 研究目标和内容

### 2.1.1 研究目标

本毕设，旨在开发一个用于匹配技术专利以及相关技术标准的算法。该算法需要基于深度学习的相关知识来找到合适的模型及其参数，同时这个模型还必须在实际的应用中——即与专利、标准相关的数据集中，取得预期中的效果。

### 2.1.2 研究内容

A 改进已有的NER算法，并通过第三方的数据集来训练合适的NER模型

在目前流行的NER算法中，选择综合了轻量、高准确度两种优点的合适的算法，并且在此基础上进行改进。同时由于项目提供的专利标准数据集并没有分词的label属性，因此无法在该数据集上训练模型，而需要通过第三方的数据集来负责训练模型，而对于模型迁移来说，不同的训练集也会影响到迁移的效果，因此如何选择合适的训练集也是一项亟待解决的问题。

B 将训练好的NER模型迁移至专利、标准数据集中，并进行匹配

在第三方的数据集上训练好所需的模型后，就需要将之迁移到本毕设真正需要处理的数据集——也即是专利、标准数据集上，我们需要对这两个数据集进行实体抽取操作，并将抽取出的实体进行匹配操作。

C 将该模型耦合进整体工程中

由于本毕设实质上是为了实现一个专利知识图谱平台上的一个子功能，因此，在毕设的最后，我还需要将毕设工作耦合进整个工程中，其中涉及前后端的对接以及相关的debug环节。

# 论文工作进展

## 2.1 总体研究进展

NER算法的研究方面，已经完成了对合适的NER数据集的选取与构建，同时在研究现有的流行NER算法的基础上，对基于深度学习的NER算法进行了优化改进，并且使用构建的NER数据集训练出了在测试集上表现相对优秀的模型。

NER模型迁移方面，已经完成了将训练好的模型移植到专利-标准数据集的工作，其中包括了对专利-标准数据集的适配工作以及参数的调整，由于专利-标准数据集并不包括分词label，因此需要人工对其采样验证准确度，我从中抽取了50个测试样本并为其人工标注比对抽取结果，发现其准确率大概与在第三方数据集上的表现吻合。

对NER处理后的专利-标准实体集间的匹配问题，目前仍然在调研阶段，初步拟采取……

对于最后的整合与DEBUG阶段，将会在所有的工作完成后进行。

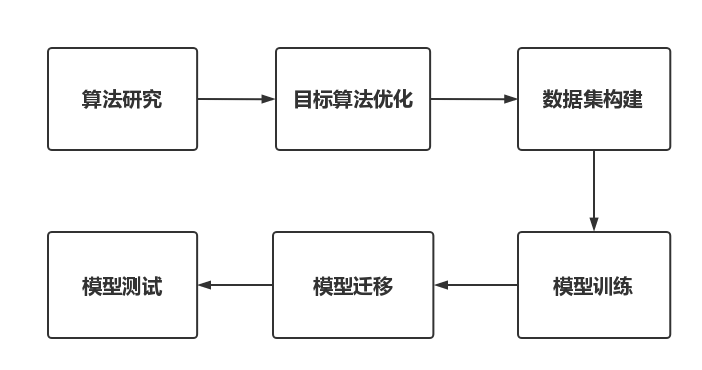
**表2.1 计划进度表**

|  |  |  |
| --- | --- | --- |
| 周 | 工作任务 | 完成情况说明 |
| 1-4 | 学习毕设所需的知识 | 完成 |
| 5-7 | 找到合适的NER算法并且在其基础上优化改进，并且在第三方数据集上训练得到相对优秀的准确率。 | 完成 |
| 8 | 将在第三方数据集上训练得到的模型迁移到标准-专利数据集上，并且调整得到适合目标数据集的最终NER模型。 | 完成 |
| 9-10 | 对NER抽取出来的专利数据实体集以及标准数据实体集间进行运用相关的深度学习方法进行匹配。 | 初步确定模型 |
| 11 | 将最终的成品模型整合进整体项目中。 | 尚未开始 |

## 

### 2.2 总体研究方案

本毕设的研究路线主要分为现有算法研究、目标算法优化、数据集构建、模型训练、模型迁移、模型测试六个阶段。流程图如图2.2。



**图2.2 研究路线**

## 2.3 主要研究成果

目前阶段，本毕设的研究成果主要有3个方面：BERT中文数据集选取与构建、BERT算法选取与优化、目标算法对应的模型训练与迁移测试。

### 2.3.1数据集选取与构建

在中文BERT研究领域，数据一直是困扰研究者的主要问题之一，数据量与颗粒度是影响BERT模型训练效果的两大主要因素。本次毕设选取的是由CLUE(中文语言理解测评基准)项目组在2020年基于清华大学的开源文本分类数据集THUCTC构建的细粒度命名实体识别数据集CLUEFineGrainNER，源数据来自Sina News Rss。

该数据集中包含训练集与验证集两部分，其中训练集10748例，验证集1343例，数据分为10个标签类别，分别为: 地址（address），书名（book），公司（company），游戏（game），政府（goverment），电影（movie），姓名（name），组织机构（organization），职位（position），景点（scene）。其中各种标签类别的具体分布如表2.3。

**表2.3 数据集标签分布表**

|  |  |  |
| --- | --- | --- |
| Label | TrainSet | TestSet |
| Address | 2829 | 364 |
| Book | 1131 | 152 |
| Company | 2897 | 366 |
| Game | 2325 | 287 |
| Government | 1797 | 244 |
| Movie | 1109 | 150 |
| Name | 3661 | 451 |
| Organization | 3075 | 344 |
| Position | 3052 | 425 |
| Scene | 1462 | 199 |
| Total | 10748 | 1343 |

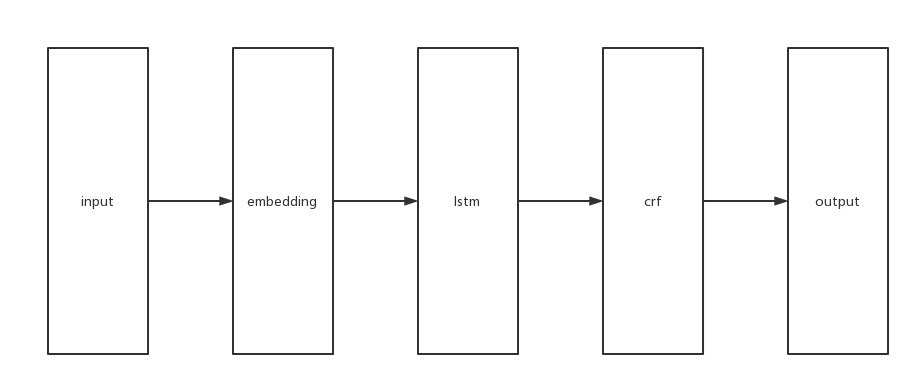
本数据集以json的格式储存，对于每条具体的数据，数据分为两个部分：text和label，其中text代表文本信息，label代表实体标签，例如：

text: "北京勘察设计协会副会长兼秘书长周荫如"

label: {"organization": {"北京勘察设计协会": [[0, 7]]}, "name": {"周荫如": [[15, 17]]}, "position": {"副会长": [[8, 10]], "秘书长": [[12, 14]]}}

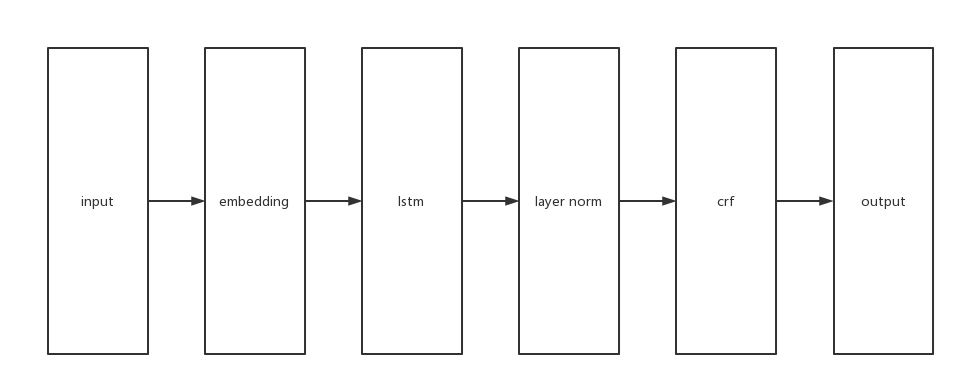
### 2.3.2 算法的选取与优化

对现有的算法的研究中，目前的流行的用于实体抽取的算法主要分为基于LSTM的各种衍生算法以及近年来的出现的基于Transformer的BERT算法。在目标算法选取上，由于本毕设的最终成果要求并非对专利-标准数据集的精确实体抽取，而是基于两个数据集各自抽取出来的实体集间进行匹配，因此个位数维度的准确性的差异并不会显著影响最终匹配结果，而BERT算法在模型训练过程中，对于算力的要求比较苛刻，在本科毕设阶段日常所能接触的算力中，都很难满足其要求。经过实测，在I7-8750H与GTX1060的环境下，同样一批数据训练同样的epoch，在改进后的LSTM算法上进行模型训练仅需45分钟，而在BERT上训练时间多达10小时。而本毕设并不是单纯的对于算法性能的学术性研究，而是需要应用于工业用途，因此在保证一定的性能的前提下，还需要尽量选取轻量级的模型算法，所以在最后我选择了基于LSTM的改进模型LSTM-CRF作为baseline。该算法的模型结构如图2.4。



**图2.3 lstm-crf模型图**

在算法优化环节，我注意到该模型为非树形结构RNN模型，因此可以在LSTM层与CRF层之间添加LayerNorm层来加速收敛速度与计算精度，优化后的模型结构为图2.5。



**图2.4 优化后的lstm-crf模型图**

### 2.3.3 模型的训练与测试

本毕设研究的基于深度学习的专利与标准关系发现技术所用的算法主要基于Pytorch深度学习框架实现，训练过程中，超参数来源于选取的baseline，而初始值则为Pytorch中随机种子函数生成。

作为在CLUEFineGrainNER数据集上模型性能的对照试验，我分别选取了由Woodbridge等人实现的不包含CRF的Lstm模型以及最新的由CLUE项目组复现并提供的BERT-base模型作为对照组，以F1与训练时长作为实验评价指标，同时排除了Other标签实体对于结果的干扰，可以得到效果对比如表2.5。

**表2.5 效果对比表**

|  |  |
| --- | --- |
| Model | F1(%) |
| LSTM | 63.62 |
| BERT-base | 78.32 |
| LSTM-CRF | 74.57 |

由该表，我们可以看到比起简单的LSTM，LSTM-CRF复合模型的效果要显著优越，同时二者的训练代价几近相同，而对比最近流行的BERT算法，由于具体应用场景中，文本的长度普遍较短，二者的性能差异不足4%，对于仅仅只是将抽取出的实体作为中间数据的本毕设而言，这些差异是在可接受范围内的，同时考虑到毕设可以接触到的算力的限制以及工业实现对于轻量级模型的需求，对于算力要求以及网络复杂度大大低于BERT模型的LSTM-CRF显然是更加适合本次毕设的模型。而对于LSTM-CRF模型，具体到每个标签（不包含Other）的测试实验效果如表2.6。其中P代表精确率(precision)，R代表召回率(recall)

**表2.5 实验具体效果表**

|  |  |  |  |
| --- | --- | --- | --- |
| Label | P(%) | R(%) | F1(%) |
| Scene | 69.7333 | 56.1159 | 62.1879 |
| Position | 80.2619 | 71.3810 | 75.5614 |
| Government | 77.9686 | 79.3313 | 78.6441 |
| Organization | 74.5660 | 67.9038 | 71.0791 |
| Name | 75.0813 | 77.6059 | 76.3228 |
| Address | 69.1335 | 64.1809 | 66.5652 |
| Game | 79.9112 | 86.8516 | 83.2370 |
| Book | 84.0561 | 63.8566 | 72.5771 |
| Company | 77.1635 | 75.7969 | 76.4741 |
| Movie | 78.5784 | 77.2989 | 77.9334 |

# 尚存的问题及后期工作安排

## 3.1 尚存的问题及措施

目前存在的问题是对于提取出来的标准与专利的实体数据集，需要找到一个合适的算法来匹配两个数据集间不同对象间的关系，同时在完成所有的工作后，还需要与负责系统开发的人员进行前后端的对接，其中可能还会涉及debug环节。为了解决这些问题，我需要继续请教老师与学长，尽快敲定下个环节实现所需的模型与算法。

## 3.2 后期工作安排

**表3.1 后期任务计划表**

|  |  |  |
| --- | --- | --- |
| 周 | 工作任务 | 完成计划（技术/手段/方法） |
| 9-10 | 对NER抽取出来的专利数据实体集以及标准数据实体集间进行运用相关的深度学习方法进行匹配。 |  |
| 11 | 将最终的成品模型整合进整体项目中。 |  |

# 

# 参考文献