单位代码 10006

学 号 18231098

分类号

密 级 公开

****

毕业设计(开题报告)

**—— 基于深度学习的专利与标准的关联关系发现技术**

|  |  |
| --- | --- |
| 院（系）名称 | 高等理工学院 |
| 专业名称 | 计算机专业 |
| 学生姓名 | 霍飞烨 |
| 学院指导教师 | 张辉 |
| 工业部门导师 |  |

2020年2月25日

# 1.论文选题依据

## 1.1研究背景

在前一段时间，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。

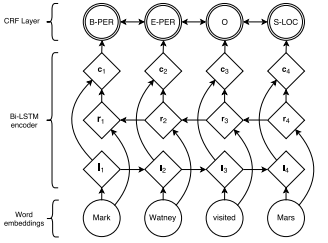
## 1.2研究意义

在深度学习，尤其是其中NLP领域的算法尚未完全成熟的过去，分析某个技术领域的专利与标准的关联，需要人工整理数据并提取其中的关系。由于效率限制，支撑结论的数据量通常不会太过庞大，且耗时巨大。而通过爬虫与深度学习相结合，则可以在短时间内大量提取特定领域的相关专利中关联的技术标准，从而更加高效准确地做出决策判断。

本毕设以服务“111”创新工程标准化需求为导向，跟踪国内外先进技术专利信息，负责对专利数据与标准数据进行关联研究，指导企业及时跟踪科技创新前沿，寻找专利申请的主要方向，开展精准创新。基于科技创新前沿专利画像模型、标签体系及标准数据，抽取专利实体的特征标签，挖掘专利、标准实体关系。

## 1.3国内外研究现状

国内外对于专利与标准的关联关系的发现这一具体的子课题的研究并不多。但是对专利与标准的关联关系的发现问题，本质上就是自然语言处理（NLP）中的命名实体识别（NER）课题。而NER是一个具有挑战性的学术课题。一方面，在大多数语言和领域，只有非常少的监督训练数据可用。另一方面，对于可以作为名称的单词种类几乎没有什么限制，因此从这个小数据样本中进行概括是困难的。因此，许多性能优秀的模型都严重依赖于人工构造的特征。为了解决这个问题，Carnegie Mellon University的Guillaume Lample、Miguel Ballesteros、Sandeep Subramanian、Kazuya Kawakami以及Chris Dyer在2016年发表的论文《Neural Architectures for Named Entity Recognition》中提出LSTM+CRF模型。它不使用任何特定于语言的资源或特性，只使用少量有监督的训练数据和未标记的语料库，是一种端到端的判别式模型。LSTM 利用过去的输入特征，CRF 利用句子级的标注信息，源序列先输入到LSTM中，然后输出一个单词的考虑到上下文的特征到CRF中，因为CRF可以考虑全局依赖。因此，该模型可以有效地使用过去和未来的标注来预测当前的标注。以序列Mark Watney visited Mars为例，其大体的模型如下：



# 2.研究方案

## 2.1研究目标及研究内容

### 1.2.1 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。

### 1.2.2 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

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。而对于迁移学习，我们很难用网络上开源数据集训练出来的模型原封不动地用于预测目标数据集，因此，我们需要手动构建自己的数据集，并用该数据集来训练已经优化好的模型算法，得到最终的深度学习模型。最后再将这个模型应用到专利-标准数据集，即可得到最后的想要的结果。

C 将毕设成果整合进项目工程中

由于本毕设实质上是山西111创新工程平台中的一个子功能，因此，在完成了本毕设后，还需要将成果整合进整个平台系统中，这就需要和负责相关模块的开发人员进行对接，并且DEBUG使得整体工程鲁棒可用。

## 2.2 拟采用的研究方法、技术路线

2.2.1 研究方法

本次毕设，将会采用文献研究法、比较研究法与实证研究法。

文献研究法：通过图书馆、互联网、电子资源数据库等途径查阅大量文献，理解前沿NER算法模型、前沿爬虫技术等相关知识，理清NER的发展脉络及研究现状，并且围绕目前调研后选择的算法模型选择性地学习深度学习有关理论知识，为设计最终的毕设成果中深度学习模型提供思路和参照。

比较研究法：在毕设调研中，除了会确定一个主要方案外，我还会通过与老师学长的交流，敲定其余一两个备选方案，当主要方案在实现过程中发现面对这个特异性问题的性能效率并未达到预期时，将会转向备选方案进行比照实验，最终选定其中相对优越的方案作为最终方案。

实证研究法：深度学习学科是一个实验的学科，再多么优越的模型都需要大量的数据集来进行训练测试，因此，本次毕设除了一开始需要用事先准备好的静态数据集进行初试模型训练与测试调整外，在工程后期，还会结合爬虫爬取的最新数据进行实战测验，经过这两轮的实验后，方会根据实验结果来决定毕设的开发是否获得预期中的效果。

2.2.2 技术路线

本次毕设开始，我首先会与老师交流明晰整个毕设的目的任务，并且进行相应的学习。首先我会学习最基础的深度学习知识，然后由于本次毕设是NLP方向相关，因此主要的学习任务将会是循环神经网络（RNN）。在学习完最基础的深度学习知识后，我将会调研目前前沿的NLP相关课题的算法模型，并且在其中挑选出符合自己能力的同时相对优越的模型作为毕设的参照模型。随后，我会根据论文的描述或者参考其源代码搭建自己的Pytorch模型，并对其不断迭代优化，当该模型在静态数据集上表现良好后，我就会利用爬虫技术来爬取动态数据进行测试直到效果达到预期。最后，我才会开发相应的前后端并封装整个程序。

## 3. 研究计划

|  |  |
| --- | --- |
| 周 | 主要任务目标 |
| 1 | 与老师学长沟通毕设成品的主要功能，初步明确接下来的学习方向，即深度学习基础知识 --> NLP方向NER算法 --> 爬虫技术 --> 前后端开发技术。 |
| 2 | 调研前沿的NER算法论文，并听取老师学长意见，从中挑选一个在我实现能力之内的作为接下来围绕学习主要方案。同时也会挑选几个作为备用方案。 |
| 3-4 | 通过吴恩达、李沐等人工智能专家出品的广受好评的教育系列视频学习深度学习的基础知识，并且选择性地学习其中与NLP相关的部分。在掌握了相关知识后，细读理解之前调研所选中的主要方案的论文，理解算法模型的原理。 |
| 5-8 | 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。 |
| 9-11 | 手动标注构建一个基于专利-标准的NER标注数据集。并将之运用至目前已经确定的模型算法中。最后，则会对抽取出专利实体数据集与标准实体数据集间进行关键词匹配。 |
| 12 | 将最终的成品模型整合进整体项目中。 |

## 4. 参考文献

1. Lample G , Ballesteros M , Subramanian S , et al. Neural Architectures for Named Entity Recognition[J]. 2016.
2. Collobert R , Weston J , Bottou L , et al. Natural Language Processing (almost) from Scratch[J]. Journal of Machine Learning Research, 2011, 12(1):2493-2537.
3. Ando R K , Zhang T . A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data[J]. Journal of Machine Learning Research, 2005, 6:1817-1853.
4. Xu L , Tong Y , Dong Q , et al. CLUENER2020: Fine-grained Named Entity Recognition Dataset and Benchmark for Chinese[J]. 2020.
5. Beryozkin G , Drori Y , Gilon O , et al. A Joint Named-Entity Recognizer for Heterogeneous Tag-sets Using a Tag Hierarchy[J]. 2019.
6. Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.
7. Wang, Xuan, Xu, et al. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN[J]. Expert Systems with Application, 2017.
8. Greff K , Srivastava R K , J Koutník, et al. LSTM: A Search Space Odyssey[J]. 2015.
9. 韩雪冬. 基于CRFs的中文分词算法研究与实现[D]. 北京邮电大学, 2010.
10. 洪铭材, 张阔, 李涓子. 基于条件随机场(CRFs)的中文词性标注方法[J]. 计算机科学, 2006, 33(10):148-151.