中文专利关键信息语料库的构建研究

张文婷[*]，赵美含[*]，马翊轩[*]，王文瑞[*]，刘宇哲[*]，杨沐昀[#]
哈尔滨工业大学计算学部，黑龙江哈尔滨150001
120L021002@stu.hit.edu.cn; yangmuyun@hit.edu.cn
邓宇
哈尔滨市阳光惠远知识产权代理有限公司，黑龙江哈尔滨150000
dy@shineip.com

摘要

专利文献是一种重要的技术文献，是知识产权强国的重要工作内容。目前专利语料库多集中于信息检索、机器翻译以及文本文分类等领域，尚缺乏更细粒度的标注，不足以支持问答、阅读理解等新形态的人工智能技术研发。本文面向专利智能分析的需要，提出了从解决问题、技术手段、效果三个角度对发明专利进行专利标注，并最终构建了包含313篇的中文专利关键信息语料库。利用命名实体识别技术对语料库关键信息进行识别和验证，表明专利关键信息的识别是不同于领域命名实体识别的更大粒度的信息抽取难题。

关键词： 专利；语料库；关键信息

Research on the construction of Chinese patent key information corpus

Wenting Zhang[*], Meihan Zhao[*], Yixuan Ma[*], Wenrui Wang[*], Yuzhe Liu[*], Muyun Yang[#]
Harbin Insititute of Technology, Harbin, Heilongjiang 150001, China
120L020815@stu.hit.edu.cn
Yu Deng
Harbin Shineip Intellectual Property Corporation, Harbin, Heilongjiang 150001, China
dy@shineip.com

Abstract

As a kind of important technology document, the patent is of substantial significance to the national intellectual property strategy in China. Existing patent corpus are mostly for the purpose of information retrieval and machine translation task, leaving the fine-grained annotated patent less touched. To facilitate the forth-coming intelligent patent technology development, this paper constructs a Patent Key Information Corpus, consisting of 313 patents annotated with the issues, methods and effects in the texts. Then the SOTA named entity recognition models are applied to the corpus, and the sharping decrease in the performance indicate the automatic identification of the key information in a patent is a challenging IE task.

Keywords: Patent , Corpus , Key Information

1 引言

[*] 排名作者不分先后，同等贡献
[#] 通讯作者

专利是专利权的简称，是由专利机构依据发明申请所颁发的一种文件 (国家知识产权局, 2008)。相对于其他文献形式，专利作为技术信息最有效的载体，不仅囊括了全球最新的技术情报，而且更具有新颖、实用、可比较、结构一致的特征。专利信息的分析利用可为企业提供技术发展路线、竞争对手动态、重点专利技术方案和技术功效矩阵，是高效开展技术攻关活动不可或缺的助手(张晓林, 2018; Balsmeier et al., 2018; 李华锋et al., 2017)。

专利被认为是世界上最大的技术信息来源。据世界知识产权组织公布的2021年度《世界知识产权指标》(WIPO, 2021)显示：2020年全球发明专利申请量达到327万件，中国发明专利申请量达150万件。现有的专利自动处理技术主要围绕检检索和翻译等文本层次粒度展开(Narin F., 1994; Mohammad Hamdaqa and Abdelwahab Hamou-Lhadj, 2009)，尚未深入探讨对专利内容的深层次细粒度的分析和理解，无法满足海量的专利数据的智能化分析及加工需求。另一方面，自然语言处理技术的近来日益成熟，信息抽取以及阅读理解应用愈发广泛。但是，当前主流的基于深度学习的自然语言处理模型往往依赖于大规模高质量的标注语料库(Sunet al., 2017)，而现有专利的语料库主要用于信息检索和跨语言翻译目的，缺乏细粒度语义标注的语料库，尚不能有效支撑智能专利处理技术的研发。

针对上述问题，本文提出聚焦专利的关键信息进行标注，提出围绕专利中的技术问题、技术方法以及技术效果这的三要素来标注该专利的关键信息。文中给出了初步的标注标准，完成了313件的专利关键信息标注语料库。并在此基础上，验证了现有主流命名实习识别模型用于自动抽取这些专利关键信息的效果，揭示了相对于日益成熟的命名实习识别任务，专利关键信息的抽取问题更具挑战性。

本文后续内容组织如下：第1节介绍目前国内外的专利语料库的相关工作，第2节介绍专利关键信息标注设计思路，第3节给出语料库构建过程，第4节对现有名实体识别技术用于专利关键信息识别进行了初步探索并给出了实验结果，第5节总结全文。

2 国内外专利语料库及概况

国外的专利领域语料库中，检索语料库和平行语料库是较为丰富的，CLEF-IP 2009 (Roda et al., 2009)、TREC-CHEM 2009 (Lupuet al., 2009)等较早时间的发布的数据集都是作为检索语料库出现的，可用于专利分类、检索等。之后出现的的SureChEMBL(Georgeet al., 2016)语料库涵盖了更多的专利文本，并且算法支持关键字检索和化学实体识别。而在平行语料库领域，典型的语料库有NTCIR6日-英双语平行语料库(Utiyama and Hi Masao, 2007)以及包含六种语言的Sentence-Aligned 欧洲专利语料库(**?**)，作者各自选取了大量专利文本，以自动标注的方式进行了句子对齐，上述语料库能为机器翻译，多语言词典等任务提供支持。

在专利领域，也出现了若干相对更为细粒度信息的标注实践。Dmitriy Korobkin等对USPTO 和RosPatent 数据库中的专利文本进行了物理效果和技术功能的标注(Korobkinet al., 2019)，；Saber A. Akhondi1等(Akhondi, S. A.et al., 2014)以手动注释为主的方式建立了一个化学专利语料库，对化合物所属种类，可合成的药物，应用目标，作用方式等做了详细注释，最终得到约76万份专利标注。

国内公开的专利语料库中，Bin Lu等(Lu, Bet al., 2009)在09年就公开过一个平行的中英文专利文本语料库，对中英文语句进行了对齐。其次还有若干检索语料库，如翟东升等(翟东升et al., 2013)对德温特专利数据库进行了信息清洗和标注，构建出了一个可用于检索分析的专利信息语料库。章成志等学者个人信息及专利成果过更加注重细粒度的手工标注，数据集主要用途是学者的画像生成(高扬et al., 2019)。

下表为本文收集整理的现有主要专利数据集的概要信息。

专利以外的各领域专业文本中，更细粒度信息的语料标注开展相对较多。李智恒(李智恒et al., 2018)人从生物医学文献中抽取化学物质致病关系的系统，崔博文(唐晓波and 刘志源, 2021)对自由文本电子病历进行了命名实体以及实体间关系的抽取，唐晓波和刘志源(唐晓波and 刘志源, 2021)针对中文金融文本领域，对重叠性较高的实体关系进行了识别，如“买卖”，“股权”，“合作”等。

当前，人工智能技术迅速发展。就自然语言处理领域来说，诸如推荐、问答等技术相对成熟，阅读理解、对话、自动写作不断取得突破。对比这些任务中所使用的标注语料的信息粒度和语义检测，以专利为代表的专业技术文本语料库建设明显存在滞后。探索建立更细粒度、更深层次能够支撑智能专利分析和理解技术研发的专利标注语料库，已经成为一项期亟需解决的

问题。

语料库/数据集	用途	语言	数据规模	标注方式
CLEF-IP 2009 Collection	现有技术搜索	英语、法语、德语	1985年至2000年100万专利和2001年至2006年50万专利	自动标注加手工标注
TREC-CHEM 2009	技术调查、现有技术搜索	各国语言	约120万份化学专利和5.9万篇科技文献	自动标注
SureChEMBL	基于关键字的检索功能	英文、德文、法文为主	1400多万份专利文件	自动标注
NTCIR6 Japanese-English Patent Parallel Corpus	机器翻译，检索	日-英平行语料（互译）	大约200万个自动对齐的句子对	自动标注为主
Sentence-Aligned European Patent Corpus	专利翻译	6种欧洲语言	1.3亿对句子	自动标注为主
a Matrix “Physical Effects –Technical Functions”	提取物理效果和技术功能	英语，俄语为主	数据库超两千万篇专利文献（未全部构建）	自动标注
Annotated Chemical Patent Corpus	分析专利内容，用于生物、化学领域的研究探索	英语为主	最终包括约76万份专利标注	手动标注
Chinese-English Patent Parallel Corpus	中英句子对齐，数据清洗	中英平行语料（互译）	16w句子对	自动标注
德温特专利数据库（DII—Derwent Innovation Index）	数据清洗，简单信息标注，构建高质量专利数据集用于专利分析和知识发现	英语	测试大小8万条专利，数据库专利一千万条以上	自动标注
杰出人才精准画像构建语料库	用户画像生成	中文	国内一亿余件专利信息	手工标注

表 1: 部分专利语料库的概要信息

3 专利关键信息的标注设计

本文旨在聚焦专利中最有价值的信息进行标注，以期支持智能专利分析技术的研发。作为阶段性成果，本节聚焦专利的关键信息进行标注，即提取技术问题、技术方法以及技术效果三类关键词来概括整篇专利，并给出了初步的标注原则。

3.1 专利文本标注需求分析

目前，针对专利的标注内容和标注粒度并没有统一的范式，针对不同的具体任务，需求各不相同。比如从专利竞争分析角度出发，专利的所属权较为关键；而从专利的行业分析角度出发，专利的所属领域更加关键。考虑到专利本身的技术文献属性，本文在首先考虑的是更为广泛和经典的技术术语标注。但是，专利的技术术语无法完整的刻画一篇专利。表2所示，其中列出的示例1和2的两个专利的术语列表，虽然大致表示了这两个专利的领域和相关技术，但是对于其专利要点以及区分这两个同主题（工业机器人）的专利来说，作用并不显著。

示例二：本发明公开了一种考虑系统延迟的不确定工业机器人运动控制方法，首先建立工业机器人机电耦合非线性动力学模型，再利用反馈线性化技术使工业机器人非线性动力学方程线性化，构建动态递归神经网络估计并补偿系统的不确定性，最后提出改进的Smith预测控制方法消除系统延迟的影响。本发明对于系统延迟和不确定性参数具有较好鲁棒性，极大地提高了工业机器人的控制精度。示例*1*：本发明公开了一种工业机器人模型仿真控制方法及装置。其中，该方法包括：接收由三维建模软件构建的工业机器人模型；基于工业机器人模型确定控制参数；根据控制参数确定工业机器人仿真机械模型；根据小脑模型神经网络*CMAC*控制策略和比例积分微分*PID*控制策略对工业机器人仿真机械模型进行仿真控制。本发明解决了相关技术中用于工业机器人的控制策略无法满足工业机器人对高速度和高精度的要求的技术问题。

示例二：本发明公开了一种考虑系统延迟的不确定工业机器人运动控制方法，首先建立工业机器人机电耦合非线性动力学模型，再利用反馈线性化技术使工业机器人非线性动力学方程线性化，构建动态递归神经网络估计并补偿系统的不确定性，最后提出改进的*Smith*预测控制方法消除系统延迟的影响。本发明对于系统延迟和不确定性参数具有较好鲁棒性，极大地提高了工业机器人的控制精度。

示例一	示例二
工业机器人	系统延迟
仿真控制	工业机器人
三维建模	运动控制
仿真机械模型	非线性动力学
小脑模型神经网络CMAC控制策略	反馈线性化技术
比例积分微分PID控制策略	动态递归神经网络
	Smith预测控制方法
	系统延迟
	鲁棒性

表 2: 专利示例中的专业术语

通过进一步调查专利相关文件的撰写要求我们发现：根据《专利审查指南》在第二部分第二章规定，专利说明书应当写明发明或者实用新型所要解决的技术问题以及解决其技术问题采用的技术方案，并对照现有技术写明发明或者实用新型的有益效果*(中华人民共和国国家知识产权局，2010)*。也就是说一个专利的关键信息包括：技术问题、技术方案以及技术效果三个部分。

进一步地，我们可以将技术问题（简称“问题”）关键词定义为专利的技术所要解决的问题，将技术方法（简称“方法”）关键词定义为解决技术问题所采用的技术方案以及关键技术手段，将技术效果（简称“效果”）关键词定义为具有技术贡献的技术方案直接带来的、或者由所述的技术特征必然产生的效果*(张晓林，2018)*。

根据这个定义，上述两个示例专利的关键信息的关键如下表所示：

我们可以发现，标注了这三个方面的关键信息之后，我们发现可以比较准确地区分出这两个专利。虽然两篇专利均为工业机器人领域，但是在问题关键词上示例一和工业机器人模型有关，而示例二和工业机器人有关且示例二考虑到了系统延迟。在方法关键词上二者所采用的方法也不同，在效果关键词上示例一提升了精度而示例二具有较好的鲁棒性，两篇专利有着实质的区别。这对于理解和梳理这一领域的专利布局、挖掘专利覆盖的方向，都具有明显的助力。

3.2 中文专利关键信息标注原则

考虑到标注的成本，本文将专利关键信息的标注范围限定在专利的标题和摘要范围。一方面可以避免下载专利全文的负担，另一方面也节省了大量的标注时间。

进一步的，我们将以以下专利的标题和摘要为例，说明我们对于问题、方法和效果这三种信息标注的适用原则。

题目：一种考虑系统延迟的不确定工业机器人运动控制方法

示例	技术问题	技术方法	技术效果
示例一	工业机器人模型仿真控制方法及装置	工业机器人模型、仿真机械模型、小脑模型神经网络CMAC控制策略、比例积分微分PID控制策略	高速度、高精度
示例二	考虑系统延迟的不确定工业机器人运动控制方法	建立工业机器人机电耦合非线性动力学模型、反馈线性化技术、动态递归神经网络、改进的Smith预测控制	鲁棒性、提高、控制精度

表 3: 专利示例中的关键信息

摘要：本发明公开了一种考虑系统延迟的不确定工业机器人运动控制方法，首先建立工业机器人机电耦合非线性动力学模型，再利用反馈线性化技术使工业机器人非线性动力学方程线性化，构建动态递归神经网络估计并补偿系统的不确定性，最后提出改进的*Smith*预测控制方法消除系统延迟的影响。本发明对于系统延迟和不确定性参数具有较好鲁棒性，极大地提高了工业机器人的控制精度。

3.2.1 技术问题关键词

在上述示例技术问题关键词为：

考虑系统延迟的不确定工业机器人运动控制方法

该关键词说明了这篇专利要解决在考虑系统延迟情况下工业机器人的运动控制方法。而技术问题关键词在实际标注过程中还可以分为两个方面，即技术问题的主体和技术问题的预期效果，分别对应上述的工业机器人和运动控制方法。技术问题关键词一般均可以直接在题目中找到，但是在一些特殊情况下如外文译为中文的专利题目中可能找不到关键词，需要从专利摘要中寻找概括。

3.2.2 技术方法关键词

上述示例的技术方法关键词为：

建立工业机器人机电耦合非线性动力学模型、反馈线性化技术、动态递归神经网络、改进的*Smith*预测控制

该关键词说明了解决系统延迟的不确定工业机器人运动控制问题所采取的具体学科知识和主要步骤，而实际标注中我们也是将对技术方法关键词的标注分为学科知识和主要步骤两大类关键词。

3.2.3 技术效果关键词

上述实例的技术效果关键词为：

较好鲁棒性、提高、控制精度 该关键词说明了上述专利提出的考虑系统延迟的不确定工业机器人运动控制领域的技术方法所取得的效果。在实际标注中，我们发现技术效果一般存在于摘要结尾，直接提取即可。

此外，我们考虑到中文的语言特点，我们在标注过程中还遵循一下标注规则：

*1)*以顿号分隔关键词

为了统一标注格式，便于后期语料库的应用，我们规定若出现多个关键词，均以顿号分隔开，并且最后一个关键词后不加标点符号，如上述技术方法关键词的提取：建立工业机器人机电耦合非线性动力学模型、反馈线性化技术、动态递归神经网络、改进的*Smith*预测控制

*2)*技术问题关键词作为一个整体短语

由于技术问题一般出现在标题或者摘要中的第一句并且为复合短语，为了保证语义的完整性，我们规定提取整个的复合短语而不将其分隔开，讲将技术问题关键词最大化，如上述技术问题关键词的提取：考虑系统延迟的不确定工业机器人运动控制方法

*3)*技术方法、效果关键词提取语义片段

由于技术方法和技术效果一般是句子内部的短语，并且动宾语之间间隔较远，若最大化提取的话会造成关键词过于冗杂，所以我们规定在动宾语距离较远的情况下单独提取动词和宾语，仅提取关键的语义片段。如上述技术关键词所示，将“提高了工业机器人的控制精度”提取为“提高、控制精度”

4 专利关键信息语料库的标注实现

本文在研究初期接受了一项实际专利分析任务：对用户提供的机器人方向的专利集合进行标注。该集合共*313*篇专利文本，本文选取其中的题目和摘要作为语料标注范围。

4.1 人工标注过程

语料库构建的核心工作是依据制定的标注规范对语料进行标注*(*訾红英*, 2020)*。由于人工智能机器人领域尚处于发展阶段，且专业性较强，而业内缺乏统一的定义和标准，为了确定与领域更加适配的标注规范和标注策略，我们将标注过程分为预标注和正式标注两个阶段，在预标注阶段采用反复标注并讨论的策略制定初步的标注规范，在正式标注阶段使用了多轮迭代标注模式进行标注规范的更新以及标注工作，如图*1*所示。

预标注有助于减少重复劳动，节省人力和资源，提高效率，提升标注的速度与精度。在预标注阶段，我们以*50*篇专利为一周期，由二名标注规范制定者分别独立进行标注，全部完成后计算一致性，并对不一致的结果进行反复分析讨论，动态更新标注规范，得出一致的标注结果。之后按照新标注规范重复该周期，直至二人一致性达到*0.85*以上，确定最终的标注规范。

在正式标注阶段，标注工作由*5*名培训筛选后的标注人员（包括规范制定者）合作完成。为提高结果可信度，采用了多轮迭代标注的策略，即：

1）将机器人专利文本随机分成五组，由五名标注人员分别标注。

2）迭代式交换标注内容，进行第二轮标注。

3）计算两次标注的一致性，对不一致的结果进行讨论，进一步完善和细化标注规范，综合前两轮结果进行第三轮标注。

4）对第三轮标注结果进行抽样检查并计算一致性，若正确率达到或超过*0.84*则认为标注结果可信。

5）最后，对仍有分歧的标注进行讨论，由规范制定人员逐一校对，修正或删除不合理项，形成机器人方向专利语料库。

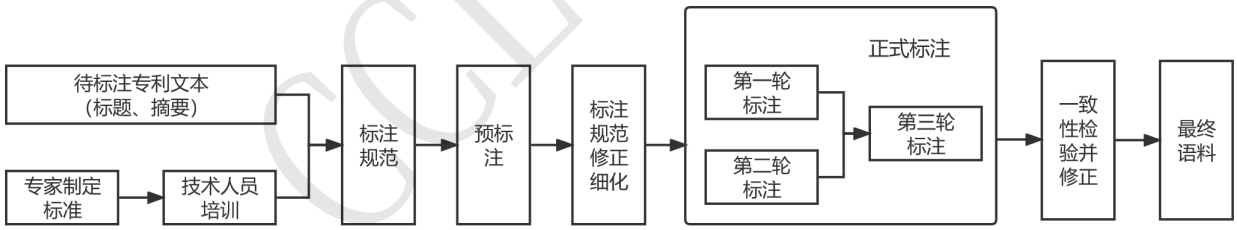

图1：语料库构建过程示意图

4.2 语料库标注质量

为了衡量语料库的标注质量，我们检验了标注结果的一致性。具体地，本文采取*Kappa*检验进行一致性检验，计算如式（*1*）所示。

$$Kappa = \frac{P_0 - P_c}{1 - P_c} \tag{1}$$

其中，P_0表示观察一致率，P_c 表示偶然一致率。我们选取五名标注人员，对全部文档进行独立标注，根据标注结果进行一致性测试。当标注术语、术语位置、术语顺序和关系类别均相同时，认为关系标注是一致的。

通过计算，本文语料标注的*Kappa*值为*0.88*，达到了用户预期的要求。进一步分析发现，由于标注人员主观认知差异，标注不一致的现象主要出现在：

1）专利问题范围的判断：不同的标注人员对同一专利所涉及的问题的范围判断不同。该类错误一般出现在标注早期，源于标注人员对标注规范不熟悉。

2）专业名词的理解：对于专利涉及的部分专业名词，不同的标注人员的理解不同，判断专利所用的技术方法时产生偏差。

专利涉及大量专业术语和专业知识，语料库的构建具有相当的挑战性。本文在标注语料库的同时也总结了一些经验：

1）进行相关名词资源的收集，其有助于界定实体边界，确定实体类型，对标注起到提示作用。

2）标注员的素质直接影响标注数据质量，因此在正式标注之前，对标注人员进行培训有利于提高准确率。

3）在正式标注过程中，对争议分歧应及时记录，定时组织讨论并听取相关专家的意见，以保证质量。

4.3 专利关键信息语料库规模统计

经过上述过程，本文最终构建了一个中文专利关键信息语料库，包含专利*313*篇，合计*9233*句、*529741*字。平均句长为*57.37*字，反映了专利文本的句子较长这一特点。

该语料库中共标注技术问题*366*个，技术方法*1384*个，技术效果*691*个。这批机器人方向专利中关注的问题主要集中在运动控制、可编程性、路径规划等；解决问题所用的技术方法主要包括深度学习、图像采集、坐标转换等；所达到的技术效果包括提高精度，提高效率，避免碰撞等。表*4*给出了三类关键词中出现频率最高的关键词及次数。

	高频关键信息(出现次数)
“技术问题”类	运动控制(47)、编程(33)、路径规划(32)、视觉(24)
“技术方法”类	深度学习(61)、图像采集(57)、坐标转换(38)、传感器(36)
“技术效果”类	提高精度(68)、提高效率(56)、加快速度(18)、避免碰撞(17)

表 4: 语料库中的各类高频关键信息

进一步地，具体到专利文本中，“技术问题”、“技术手段”、“技术效果”的关键词平均个数分别为*1.30*、*4.42*和*2.17*。进一步分析发现，“技术问题”“技术手段”“技术效果”关键词在标注文本中通常先后依次出现，第一次出现位置分别集中在标注文本的%*0.15*、%*0.2*、%*2.0*处。这一规律为我们日后进行自动化的专利阅读的深入研究提供帮助。

5 基于NER模型的专利关键信息的识别

分析上文实现的语料库的技术问题、技术方法、技术效果的标注结果，我们不难发现大量的技术专有名词（术语），一个很直接的想法就是使用成熟的命名实体识别技术对专利语料库关键信息进行自动识别。

具体地，本文用了*ACL 2021*和*ACL 2020*所发表的两个具有代表性的命名实体识别模型,分别为*Mect (Shuang Wu et al., 2021)*和*Flat Lattice Transformer (Xiaonan Li et al., 2020)*。其中，*MECT*模型将字特征、词特征和部首特征结合并能够使用多元数据特征，且在多个中文*NER*数据集上性能表现出色。*Flat Lattice Transformer*模型中所有字符都可以与其自匹配词直接交互，并可以对远距离依赖进行完全建模，且在多个*NER*数据集上被验证优于基线模型和其他基于词典的模型。同时，本文也提供了更为经典的基准*CRF*模型进行实验。

实验中，我们将*313*项专利按照*8*：*1*：*1*的比例，随机选择*251*项作为训练集，*31*项作为开发集，*31*项作为测试集。评测采用传统命名实体识别的召回率*(R)*、准确率*(P)*和F *1*值，各模型的性能指标如表*5*所示。

$$P = \frac{\text{预测正确的实体数}}{\text{预测的实体总数}} \tag{2}$$

$$P = \frac{\text{预测正确的实体数}}{\text{标注的实体总数}} \tag{3}$$

$$R = \frac{2 * P * R}{P + R} \tag{4}$$

模型	F1	准确率	召回率
Mect	0.371	0.457	0.312
Flat Lattice Transformer	**0.474**	0.512	0.440
CRF	0.394	0.558	0.305

表 5: 语料库中的各类高频关键信息

从表中可以看出，*Flat Lattice Transformer*在所使用的三个模型中性能表现最好，但是*0.474*的*F1*值远远不能令人满意。以在*NER*中的*MSRA*数据集为例，*Flat Lattice Transformer*的*F1*值为*94.12*，远高于本节的实验结果。

进一步对比本文语料库和*NER*语料我们发现:

（*1*）*MSRA*数据集有超过*5* 万条中文命名实体识别标注，本文标注的语料库只有*2441*条标注，数据规模相对较小;

（*2*）本文语料中问题关键词平均词长*13.45*字，方法关键词平均词长*6.46*字，效果关键词平均词长*8.26*字，这些都远远超过了一般命名实体的长度。

实验结果反映出本文所标注的专利关键信息自动识别是一个有待深入探索的难题，仅仅沿用命名实体模型不能充分的满足专利的信息识别需求。同时由于当前标注语料规模远远小于*NER*中的可用资源，小样本学习问题将使专利关键信息的识别，更具有挑战性。

6 结论与展望

本文面向专利智能分析技术研发的需要，研制了一个专利关键信息语料库，完成了语料库的标注设计以及实际标注，初期完成的语料库中包含*313*篇专利，*9233*个句子以及*2441*条标注。

在此基础上，本文验证了当前主流的命名实体识别模型用于专利关键信息自动识别的效果。实验标明，相对于*90%*以上的命名实体识别效果，这些模型在本文的数据上*F1*值最好只能达到*0.474*，说明专利的关键信息识别是一个有待深入探索的问题。

目前专利文本的细粒度标注的研究还比较少，本文的专利来源，题材和覆盖面相对有限，只是这方面的一个初步尝试。下一步将扩大数据规模和丰富语料来源，进一步完善标注体系，为领域知识库的构建奠定基础，同时将探索在专利语料库上有效进行关键信息识别的模型和方法。

参考文献

国家知识产权局. *2008.* 专利. *www.cnipa.gov.cn/art/2008/4/3/art_2147_152059.html*

张晓林 *2018.* 专利技术情报分析模型构建及其应用研究. 图书馆杂志,第*10*期,*78-88*

Benjamin Balsmeier, Mohamad Assaf, Tyler Chesebro. 2018. Machine learning and natural language processing on the patent corpus: data, tools, and new measures. *Journal of Economic & Management Strategy,*第*27*卷,*535-553*

李华锋，袁勤俭，陆佳莹等. *2017.* 国内外专利情报分析方法研究述评. 情报理论与实践,第六期,*139-144*

World Intellectual Property Organization. 2021. World Intellectual Property Indicators 2021.

Narin F. 1994. Patent bibliometrics *Scientometrics,*第*30*期,*147-155*

Mohammad Hamdaqa and Abdelwahab Hamou-Lhadj. 2009. Citation Analysis: An Approach for Facilitating the Understanding and the Analysis of Regulatory Compliance Documents. *Proceedings of Sixth International Conference on Information Technology: New Generations. Las Vegas, Nevada: 27-29*

*C. Sun, A. Shrivastava, S. Sing*等. *2017.* Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. *Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV),2017,pp: 843-852*

Roda, Giovanna , Tait, John , Piroi, Florina , Zenz, Veronika. 2009. CLEF-IP 2009: retrieval experiments in the intellectual property domain. *Cross-Language Evaluation Forum*

Lupu, M. , Huang, J. , Zhu, J. , Tait, J. 2009. TREC-CHEM: large scale chemical information retrieval evaluation at TREC. *Acm Sigir Forum ,43,pp:63-70*

Utiyama, Hi Masao. 2007. A Japanese-English patent parallel corpus. *Proc Mt Summit XI,2007*

George, Papadatos , Mark, Davies , Nathan, Dedman , Jon, Chambers , Anna, Gaulton , James, Siddle , Richard, Koks , Irvine, Sean A. , Joe, Pettersson , Nicko, Goncharoff 2016. SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic Acids Research,D1,2016,pp:D1220-D1228*

Tger, Wolfgang 2011. The Sentence-Aligned European Patent Corpus. *Proceedings of the 15th Annual conference of the European Association for Machine Translation,2011*

Korobkin, Dmitriy , Shabanov, Dmitriy , Fomenkov, Sergey , Golovanchikov, Alexander. 2019. Construction of a Matrix "Physical Effects – Technical Functions" on the Base of Patent Corpus Analysis.

Akhondi, S. A. , Klenner, A. G. , Tyrchan, C. , Manchala, A. K. , Boppana, K. , Lowe, D. , Zimmermann, M. , Jagarlapudi, Sarp , Sayle, R. , Kors, J. A. 2014. Annotated Chemical Patent Corpus: A Gold Standard for Text Mining. *PLoS ONE,9,9,pp:e107477*

Lu, B. , Tsou, B. K. , Zhu, J. , Tao, J. , Kwong, O. Y. 2009. The Construction of a Chinese-English Patent Parallel Corpus.

翟东升, 李倩, 张杰, 黄鲁成, 赵京 *2013.* 德温特专利信息清洗与标注模型研究. 情报杂志,*32,8,pp:6*

高扬, 池雪花, 章成志, 孔捷 *2019.* 杰出人才精准画像构建研究——以智能制造领域为例. 图书馆论坛,*39,6,pp:8*

李智恒, 桂颖溢, 杨志豪, 林鸿飞, 王健. 基于生物医学文献的化学物质致病关系抽取. 计算机研究与发展, *55*, *1*, *pp:9*

崔博文, 金涛, 王建民. 自由文本电子病历信息抽取综述. 计算机应用,*2021*

唐晓波, 刘志源. 金融领域文本序列标注与实体关系联合抽取研究. 情报科学,*39,5,9,*

中华人民共和国国家知识产权局. *2010.* 专利审查指南. 北京：知识产权出版社, *2010:2-13*

朱宝华. *2019.* 浅谈如何撰写高质量专利申请文件. 中国发明与专利,*2019,16(03):95-100*

眷红英 *2020.* 面向儿科疾病的命名实体及实体关系标注语料库构建及应用. 中文信息学报*(05),19-26*

Shuang Wu et al. 2021. MECT: Multi-Metadata Embedding based Cross-Transformer for Chinese Named Entity Recognition.. *Proceedings of the Annual Meeting of the Association for Computational Linguistics,2021:1529-1539.*

Xiaonan Li et al. 2020. FLAT: Chinese NER Using Flat-Lattice Transformer.. *Proceedings of the Annual Meeting of the Association for Computational Linguistics,2020:6836-6842.*