计算机科学
COMPUTER SCIENCE

军事指控保障领域命名实体识别语料库的构建

杜晓明，袁清波，杨帆，姚奕，蒋祥

引用本文

杜晓明，袁清波，杨帆，姚奕，蒋祥. 军事指控保障领域命名实体识别语料库的构建[J]. 计算机科学, 2022, 49(6A): 133-139.

DU Xiao-ming, YUAN Qing-bo, YANG Fan, YAO Yi, JIANG Xiang. Construction of Named Entity Recognition Corpus in Field of Military Command and Control Support[J]. Computer Science, 2022, 49(6A): 133-139.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210400132

军事指控保障领域命名实体识别语料库的构建

杜晓明　袁清波　杨　帆　姚　奕　蒋　祥
陆军工程大学指挥控制工程学院　南京 210007
(bdar@163.com)

摘　要　军事指控保障领域知识图谱的构建是军队信息化装备保障过程中的一个重要研究方向。针对保障领域知识图谱构建中命名实体识别模型缺乏相应基础训练语料库的现状,在分析相关研究现状的基础上,设计并实现了一个基于 PyQt5 应用程序基本框架的 GUI 命名实体识别语料库构建系统。首先,简要说明了系统整体架构和语料处理技术流程;其次,详细介绍了系统的数据预处理、标注体系、自动标注、标注分析和编码转换五大功能模块的相关内容,其中自动标注功能模块中的自动标注和自动去重算法的实现是重难点,也是整个系统的核心;最后,通过 PyQt5 应用程序基本框架和各类功能组件对各功能模块进行了图形用户界面实现。本系统的设计与实现,可以在军队专用电脑上对各种原始装备手册进行自动化处理,快速生成命名实体识别模型训练所需语料库,从而为后续构建相应领域知识图谱提供有效技术支持。

关键词:命名实体识别;语料库;自动标注;军事指控保障;知识图谱
中图法分类号　TP391

Construction of Named Entity Recognition Corpus in Field of Military Command and Control Support

DU Xiao-ming,YUAN Qing-bo,YANG Fan,YAO Yi and JIANG Xiang
College of Command and Control Engineering,Army Engineering University of PLA,Nanjing 210007,China

Abstract　The construction of the knowledge graph in the field of military command and control support is an important research direction in the process of the military information equipment support. Aiming at the current situation that the named entity recognition model lacks the corresponding basic training corpus in the construction of the guarantee domain knowledge graph,based on the analysis of the relevant research status,this paper designs and implements a GUI named entity recognition corpus construction system based on the basic framework of the PyQt5 application program. First,it briefly describes the overall system architecture and corpus processing technical process. Secondly,it introduces the system's data preprocessing,labeling system,automatic labeling,labeling analysis and coding conversion related content in five major functional modules. Among them,the automatic labeling function module is automatic. The implementation of automatic labeling and the realization of automatic de-duplication algorithm is the most important and difficult point,and also is the core of the entire system. Finally,the graphical user interface of each functional module is implemented through the basic framework of the PyQt5 application program and various functional components. The design and implementation of this system can automatically process various original equipment manuals on military computers,and quickly generate the corpus required for named entity recognition model training,so as to provide effective technical support for the subsequent construction of the corresponding domain knowledge graph.

Keywords　Named entity recognition,Corpus,Automatic annotation,Military command and control support,Knowledge graph

1　引言

指挥控制(Command and Control,C2)简称指控,是指参谋人员利用设备、器材,通过对信息进行收集、传输、处理和利用,为指挥员提供作战辅助决策的一门科学[1]。近年来,随着军队信息化建设步伐的加快和数字化技术的快速发展,大量新型指挥控制装备陆续配发到各级部队,这些装备是典型的技术密集型装备,具有更新周期短、体系运用要求高、人机结合紧密、运用保障难度大等特点,它们在给部队提升作战指挥能力的同时,也给部队的信息保障工作提出了新的要求。

当前,围绕对指挥控制系统的保障,部队一方面积累了大量的技术手册、操作规程、培训讲义、实施记录、保障案例等数据资料。另一方面,由于这些数据资料散落各处、形态各异、详略不一,部队在使用上效率低下、效益不高,形成了面对海量保障数据,技术人员却无从获取有用知识,也无法依据数据进行更精准、有效的保障决策的尴尬局面。因此,如何对这些数据资料进行挖掘处理和高效融合应用,已成为当前部队指控保障领域急需解决的重要问题之一。

知识图谱(Knowledge Graph,KG)技术在进行关键数据获取、有效信息融合、知识驱动应用等方面展现出巨大优势,

基金项目:全军军事类研究生资助课题(JY2019C078)
This work was supported by the Military Postgraduate Funding Projects of the PLA(JY2019C078).
通信作者:袁清波(12661967@qq.com)

已成为当前及未来知识及大数据应用领域的一个重要研究方向[2]。命名实体识别(Named Entity Recognition,NER)则主要从多源异构数据文本中识别出特定类型的实体,是自动化构建知识图谱过程中的一项重要基础性工作。然而,军事领域由于数据专业性、保密性、获取难度大等特殊原因,在命名实体识别方面进行研究的人员相对较少,相关基础性工作和技术发展较为缓慢。一方面,由于军事领域的特殊性,能够通过互联网等公开渠道来获取的支撑命名实体识别模型训练的相关语料库较少;另一方面,直接将通用领域的命名实体识别语料库移植到军事领域中,效果不佳,远远达不到应用程度的要求。

鉴于此,本文将在分析近年命名实体识别语料库构建相关研究现状的基础上,研发针对军事指控保障领域的命名实体识别语料库构建系统,从而为后续军事指控保障领域命名实体识别模型和训练语料的构建研究提供技术支持。

2 相关研究现状

近年来,知识图谱构建及其相关技术的研究正在呈增长趋势,越来越多研究者开始进入到该领域。其中,命名实体识别研究,不管在通用领域,还是在特定领域,都是知识图谱构建中的研究热点。从方法技术上看,命名实体识别研究正由基于传统的机器学习方法(如 CRF)向基于深度学习的方法(如 LSTM)发展,基于深度学习的命名实体识别方法已占据了主导地位,且取得了不错的效果。最新研究出现了结合注意力机制及 Bert 等预训练词向量模型的方法,实验效果越来越好。从研究领域上看,命名实体识别技术不仅涉及通用领域,而且覆盖医学、交通、军事等多个垂直领域。因此,构建具体适用于各领域的命名实体识别语料库势在必行。

中文命名实体识别语料库构建的相关研究,起初主要集中于通用领域。通用领域命名实体识别语料库,规模较大且质量较高的主要有 1998 年《人民日报》语料库 PKU、微软亚洲研究院语料库 MSR 等,这些都主要是基于开放互联网上的新闻领域语料而构建。此外,Li 等[3]以已有维基百科中文命名实体语料库中的标注信息为基础,自动地构造出了尽可能多的嵌套命名实体,再进行手工调整,从而构建出了高质量的中文嵌套命名实体识别语料库。

特定领域命名实体识别语料库构建研究主要集中于医学、交通、军事等领域。Yang 等[4]面向医学领域,以 992 份中文电子病历文本为基础,结合中文电子病历的特点,提出适合中文电子病历的命名实体标注体系,制定了命名实体详细标注规范,构建了中文电子病历命名实体标注语料库。Zou 等[5]面向儿科疾病领域,制定了适合儿科医学的命名实体标注体系及详细标注规范,对 298 余万字的儿科医学文本中的实体进行了标注校对,构建了包含 504 种儿科常见疾病的儿科疾病领域医学实体语料库。Mo 等[6]面向公路桥梁定期检测领域,以 150 份真实桥梁检测报告文本作为标注语料,定义了由桥梁实体、结构实体、结构病害实体等 6 种目标命名实体类别及其标注规范,构建了一个较大规模的公路桥梁定期检测领域命名实体识别语料库。Zhou 等[7]面向军事领域,针对军事语料中存在的实体,提出了一套统一的军语词性标记规范和军事语料标注规范,设计了一种基于军语词典的自动扩展的军事语料实体特征提取框架,构建了一个较大规模的高质量军事语料库。Feng 等[8]面向国防科技领域,以 479 篇文章军事技术文本为基础,在分析文本特点的基础上制定了一系列标注规范并进行相应标注工作,构建了面向国防科技领域的技术和术语语料库。

具体到军事指控保障领域,由于装备保障手册文本资料的特殊性,文档内容方面和实体类型方面都极具特殊性,以上各领域语料库构建中所用到的方法和手段虽然具有一定借鉴意义,但是都不能完全直接移植到该领域中。因此,针对军事指控保障领域的特点,以原始装备技术手册文本作为输入,以语料标注文件作为输出,来构建一套适合于该领域特点的命名实体识别语料库系统是很有必要的。

3 系统总体架构

3.1 系统架构

军事指控保障领域命名实体识别语料库构建系统总体分为 3 层,从下到上分别为数据层、功能层和表示层。体系架构如图 1 所示[9]。

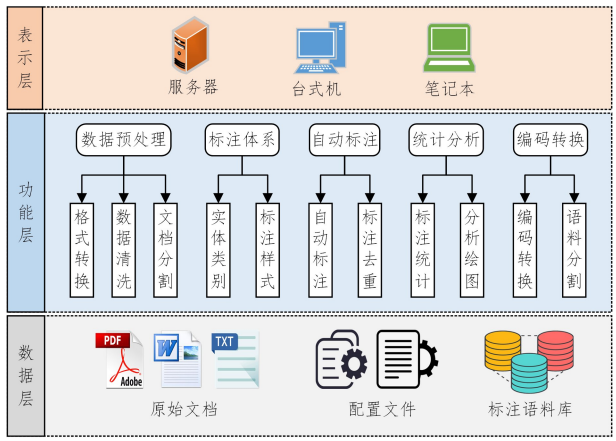

图 1 命名实体识别语料库构建系统体系架构图

Fig. 1 Architecture diagram of building system of named entity recognition corpus

数据层:主要完成系统底层所用到的各类数据文档的存储。数据文件主要分为 3 类:1)各类装备手册资料原始文档,主要有 pdf,doc,docx 以及 txt 等格式类型,作为系统输入使用;2)各类配置文件,包括标注系统所用到的定义实体类别、标注样式以及领域字典等文件;3)语料标注文件,主要是系统所输出的文件,记录着实体位置、实体类别等信息。

功能层:整个系统的核心部分,定义了系统所实现的数据预处理、标注体系、自动标注、统计分析以及编码转换五大功能模块。数据预处理功能模块主要完成文档格式转换、文本数据清洗以及分割文档成多个样本集等功能;标注体系功能模块主要完成命名实体类别定义、标注样式定义等功能;自动标注功能模块主要完成领域词典制作、自动分类标注以及标注结果去重等功能;统计分析功能模块主要完成标注语料中各类实体数量的统计以及分析绘图功能;编码转换功能模块主要完成命名实体识别自动标注文件到 BIEO 或 BIO 编码语料库的格式转换,以及将标注文件分类成用于命名实体识别模型可用的训练集、验证集和测试集。其中,自动标注功能模块是系统实现过程中的重点和难点。

表示层:整个系统的用户交互部分,通过桌面图形用户界面进行展现,其中使用到了 PyQt5 中的窗体可视化设计、

界面组件使用、事件处理、文件、绘图等主要技术。根据上述功能层所定义的各功能模块，设计出合理美观的系统界面窗体布局，方便用户进行操作使用是表示层需要关注的重点。

3.2 技术流程

基于命名实体识别语料库构建系统的体系架构，所设计的语料库构建技术流程如图 2 所示。首先，系统读取原始手册文档：1)通过数据预处理生成待标注语料文件；2)通过对文档分析和领域本体建模，完成标注实体类别的定义；3)对文档进行关键字符统计分析，形成领域词典文件。其次，在上述工作的基础上，通过开发的自动标注程序完成语料标注，形成已标注语料文件。然后，通过自动编码格式转换将已标注语料文件转换成命名实体识别模型可以训练使用的语料库。最后，对标注语料文件进行统计与分析，形成统计分析数据文件和图像文件。

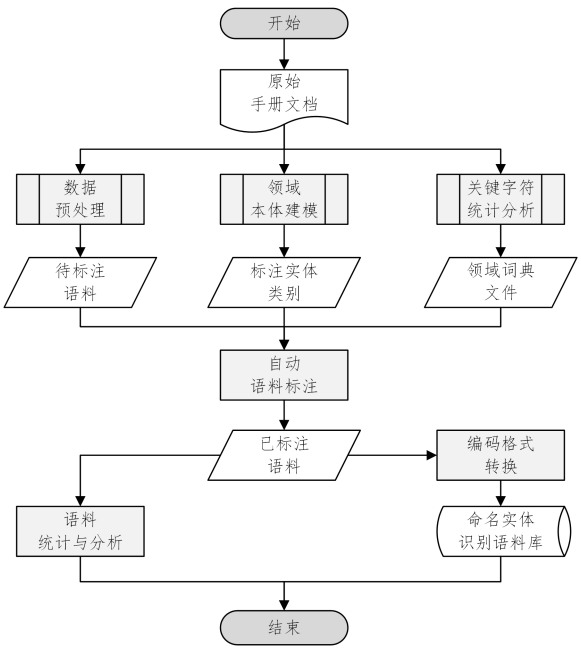

图 2 命名实体识别语料库构建技术流程图

Fig. 2 Flow chart of building named entity recognition corpus

4 系统功能模块设计

4.1 数据预处理模块

数据预处理功能模块主要完成文档格式转换、文本数据清洗以及文档分割等功能。具体实现流程如图 3 所示。

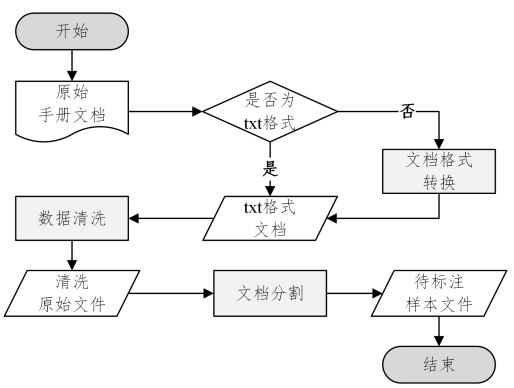

图 3 数据预处理功能模块的实现流程

Fig. 3 Implementation process of function module of data preprocessing

(1)文档格式转换。军事指控保障领域电子装备手册资料的原始文档类型多样，常见的主要有 pdf，doc 及 docx 等格式。系统在进行处理时，无法对这些格式类型的原始文档直接进行处理，首先需利用专用程序将这些格式类型文件统一转化为 txt 纯文本格式类型文件。

(2)文本数据清洗。程序转化后的 txt 格式类型文档中存在着大量的噪声数据，通常是一些无意义的特殊字符，如各类空白字符、冗余标点符号、空白行等，因此需对这些噪声数据进行去除。

(3)文档分割。文本数据经过清洗后，为便于后续任务处理，另外一项重要的工作就是对文档进行句子级别划分，使文档中每一个句子占一行位置。按照一定的句子数量(如 100 句)对文档进行分割，形成多个待标注的样本文件。常见中文句子分割标识符主要有：句号(。)、叹号(！)、问号(？)、句号＋双引号(。”)、叹号＋双引号(！”)以及问号＋双引号(？”)。

4.2 标注体系模块

标注体系功能模块主要完成实体类别及标注样式定义功能。具体实现流程如图 4 所示。

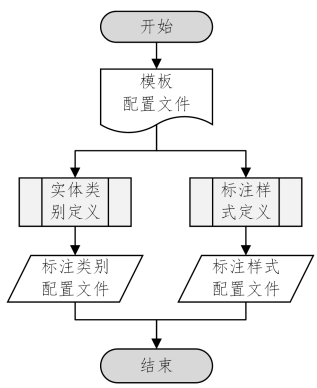

图 4 标注体系功能模块的实现流程

Fig. 4 Implementation process of function module of annotation system

(1)实体类别定义。通过分析军事指控保障领域装备资料，结合领域专家指导意见，对指控保障领域进行了本体构建，具体如图 5 所示。将指控保障领域的实体划分为了两个层次，第一层次实体共含有 5 个大类，分别为设备软件类、功能参数类、人员机构类、故障错误类和检查维修类；第二层次实体共含有 12 个小类，分别为指控车辆类、指控设备类、指控软件类、装备功能类、装备参数类、人员用户类、组织机构类、故障现象类、故障类型类、故障原因类、检查方法类和维修程序类。根据实际应用情况，这里选择将第一层次的 5 个大类实体类别定义为语料库构建过程中的实体类别。

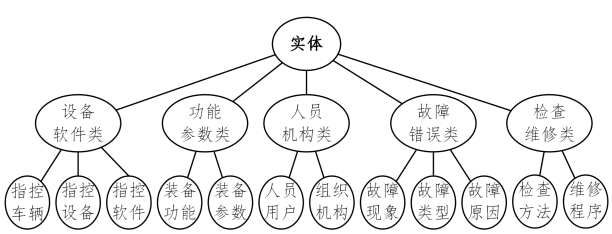

图 5 指控保障领域的本体定义

Fig. 5 Ontology definition of command and control support domain

(2)标注样式定义。标注样式指的是在标注系统中各类型

实体显示的字体颜色、背景色，以及定义的实体类别的英文及其缩写形式。颜色的设置采用十六进制表示形式，如红色为FF0000，绿色为00FF00。实体类别的英文及其缩写形式为：设备软件类（Equipment，EQU）、功能参数类（Function，FUN）、人员机构类（Organization，ORG）、故障错误类（Fault，FAU）和检查维修类（Maintenance，MAI）。其中，英文缩写部分对应着英文全称的前3个字母。定义好英文缩写后，可以便于后续编码转换部分使用。

4.3 自动标注模块

自动标注功能模块主要完成自动分类标注与标注结果去重功能，是系统实现过程中的重点和难点。具体实现流程如图6所示。

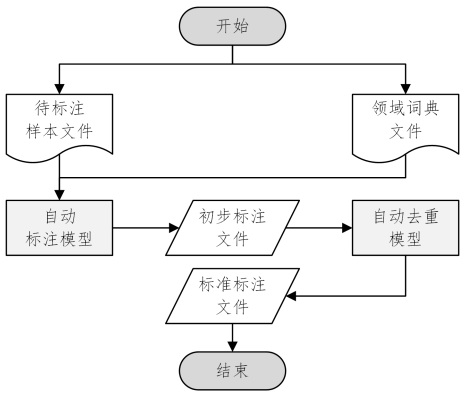

图6 自动标注功能模块的实现流程

Fig. 6 Implementation process of function module of automatic labeling

（1）领域词典文件。指控保障领域实体术语类别词典由两部分组成："1）实体术语词条＋2）实体类别"。第1部分为实体术语词条，是指控保障领域术语词典中的词条；第2部分为该术语词条的实体类别，具体类别由命名实体识别标注规范定义。

以设备软件类实体术语词典文件为例，其词典格式如表1所列。

表1 实体术语类别词典

Table 1 Dictionary of entity term category

实体术语词条	实体类别
车载计算机	设备软件类
操作系统	设备软件类
—	—

（2）输出标注文件。标注文件采用开源Brat手动标注软件可识别的ann标注文件格式[10]。ann标注文件由5部分组成：1）序号，由程序自动进行编号；2）实体类别，具体由命名实体识别标注规范所定义；3）实体术语起始位置，指待标注的术语词条在原始文件中所处的字符起始位置，文件最开始位置为0，每一个字符（不区分中英文字符、标点符号及特殊字符）占一个位置；4）实体术语结束位置，指待标注的术语词条在原始文件中所处的字符结束位置（不包含此位置）；5）具体实体，也就是待标注的术语词条。

例如，有原始文本"车载计算机操作系统无法启动、报错。"。自动标注模型输出的标准标注文件结果如表2所列。

表2 标注文件内容

Table 2 Contents of annotation file

序号	实体类别	起始位置	结束位置	具体实体术语
T1	设备软件类	0	5	车载计算机
T2	设备软件类	5	9	操作系统
—	—	—	—	—

（3）自动标注模型。自动标注模型的输入为原始语料文本和实体术语类别词典，输出为Brat标注软件可识别的ann标注文件。自动标注模型算法实现的伪代码如算法1所示。

算法1 自动标注算法

```
输入：rawCorpusFile；entityClassDict
输出：rawAnnFile
1. rawCorpusString ← Read(rawCorpusFile)
2. entityClassList ← Read(entityClassDict)
3. sequenceNumber ← 0
4. labeledEntityList ← []
5. for eachEnityClass in entityClassList do
6.     entityName ← eachEnityClass[0]
7.     className ← eachEnityClass[1]
8.     entityLenth ← Len(entityName)
9.     stringLenth ← Len(rawCorpusString)
10.    i ← 0
11.    While entityName in rawCorpusString[i:]
12.        tempList ← []
13.        tempList.append(sequenceNumber)
14.        sequenceNumber ← sequenceNumber+1
15.        tempList.append(className)
16.        startLablePosion ← rawCorpusString.index(entityName,i,
           stringLenth)
17.        tempList.append(startLablePosion)
18.        endLablePosition ← startPosion+entityLenth
19.        tempList.append(endLablePosition)
20.        tempList.append(entityName)
21.        i←endLablePosition
22.        labeledEntityList.append(tempList)
23.    end
24. end
25. for element in labeledEntityList do
26.    rawAnnFile ← Write(element)
27. end
```

自动标注具体过程如下所示：

1）读取原始语料文本，将读取到的内容保存为一个整字符串，其中每一行的换行符"\n"也占一个字符位置；

2）按行读取实体术语类别词典，将读取到的内容保存为一个列表，其中每一行的实体术语及其类别为列表的一项内容；

3）遍历词典列表，获取每一个实体术语词条在原始语料文本字符串中的所有位置，包括起始位置和终止位置，加入序号、实体种类以及实体术语后形成一个语料标注列表；

4）将语料标注列表保存到Brat标注软件可识别的ann文件中，文件中的一行为一个标注实体列表项内容，特别注意要明确ann文件的换行符为"\n"格式。

其中，获取字符串的位置是整个标注算法实现的核心和重点。

（4）标注去重模型。标注去重模型算法的输入为自动标注模型生成的标准标注文件，输出为去重后的标注文件。标注去重模型算法伪代码如算法2所示。

算法 2 标注去重算法

```
输入：rawAnnFile
输出：newAnnFile
1. labeledEntityList ← Read(rawAnnFile)
2. sortedLabeledEntityList ← Sort(labeledEntityList)
3. l← Len(sortedLabeledEntityList)
4. for i←0 to l−1 do
5.     if i=l−1 then
6.         break
7.     end
8.     currentEntity ← sortedLabeledEntityList[i]
9.     for nextEntity in sortedLabeledEntityList[i+1:] do
10.        if int(nextEntity [2]) < int(currentEntity [3]) then
11.            sortedLabeledEntityList. Remove(nextEntity)
12.        else
13.            break
14.    end
15.    end
16. end
17. for element in sortedLabeledEntityList do
18.    newAnnFile ← Write(element)
19. end
```

标注去重的具体过程如下所示：

1）按行读取 ann 标注文件，将读取到的内容保存为一个列表，其中每一行的标注实体为列表的一项内容。

2）对标注实体列表进行排序，对列表中所有标注实体，首先按照起始位置进行升序排序，起始位置数字较小的排在最前面；起始位置相同的，再按照结束位置进行降序排序，结束位置数字较大的排在最前面。

3）对排序列表中相邻标注实体间的位置关系进行比较。如果后一个标注实体的起始位置小于前一个标注实体的结束位置，说明两个标注实体间存在重叠标注，则将后一个标注实体删除；否则，说明两个标注实体间不存在重叠标注问题。循环比较，直至到达最后一个标注实体，则证明已全部比较完毕。

4）将去重后的列表保存到新的 ann 文件中，其中文件中的一行为一个标注实体列表项内容，特别注意要明确 ann 文件的换行符为“\n”格式。

4.4 标注分析模块

标注分析功能模块主要是完成标注语料中各类实体数量的统计以及分析绘图功能。具体实现流程如图 7 所示。

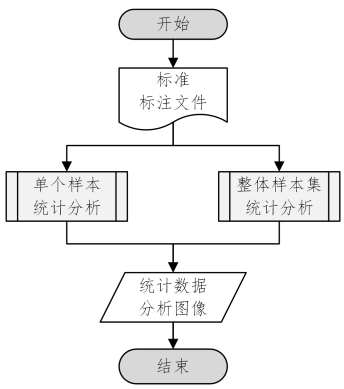

图 7 标注分析功能模块的实现流程

Fig. 7 Implementation process of function module of annotation analysis

（1）实体数量统计。原始文件标注完成后，由于标注文件数量多，且单个文件内各类型标注实体数量不一，为快速了解整体标注情况，需要对自动标注生成的标注文件进行统计分析。数量统计分为单个标注文件的统计与整个目录内所有标注文件的整体汇总统计。

（2）数据分析绘图。分析绘图功能是对统计结果的可视化展示，可以让用户更直观地了解各类型实体数量对比情况。同样，分析绘图功能也分为对单个标注文件的数据分析绘图与对整个目录内所有标注文件整体进行数据分析绘图。

4.5 编码转换模块

编码转换功能模块主要完成命名实体识别自动标注文件到 BIEO 或 BIO 编码语料库的格式转换，以及将标注文件分类成用于命名实体识别模型可用的训练集、验证集和测试集。具体实现流程如图 8 所示。

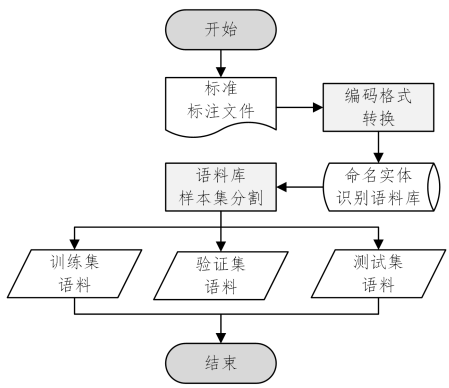

图 8 编码转换功能模块的实现流程

Fig. 8 Implementation process of function module of code conversion

（1）编码格式转换。命名实体识别本质上是对文本中的词语进行状态标注，并对实体类别进行预测。在命名实体识别语料标注时，常采用 BIEO 编码格式，其中“B”(Begin)表示当前字符为实体词语的开始字符；“I”(Inside)表示当前字符为实体词语的内部字符；“E”(End)表示当前字符为实体词语的结束字符；“O”(Outside)表示当前字符为可识别的实体词语以外的其他字符。BIEO 编码格式文件中记录了每一个字符的标识信息，可以作为命名实体模型的训练输入数据文件。

指控保障领域不同类别的实体标注方式如表 3 所列。

表 3 不同类别实体的标注方式

Table 3 Annotation methods of different types of entities

实体类别	开始标识	内部标识	结束标识
设备软件类	B-EQU	I-EQU	E-EQU
功能参数类	B-FUN	I-FUN	E-FUN
人员机构类	B-ORG	B-ORG	E-ORG
故障错误类	B-FAU	I-FAU	E-FAU
检查维修类	B-MAI	I-MAI	E-MAI

实体标注方式中所用到的“EQU”“FUN”“ORG”“FAU”和“MAI”为英文单词 Equipment，Function，Organization，Fault 和 Maintenance 的缩写形式。所有类型实体以外的字符均用“O”进行标注。

（2）样本集分割。为了便于制作的命名实体识别语料库能够直接用于后续模型训练和检验，这里将语料库按照 70%，10%，20%的比例分为训练集、验证集和测试集。其中，训练集主要用于模型的训练；验证集主要用于调试模型参数；防止模型过拟合；测试集用于最终的模型测试。

5 系统用户界面实现

军事指挥控制保障领域命名实体识别语料库构建系统，用户界面实现部分基于 Python 编程环境开发，可视化窗体使用 Qt Designer 进行设计，采用了 PyQt5 应用程序基本框架和各类功能组件。该系统采用图形用户界面(Graphical User Interface，GUI)应用程序进行开发的主要考虑是，军事指挥保障领域的原始资料文档大都具有密级，需要在部队专用电脑内进行处理，而这些电脑大都不允许接入网络，只能单机运行，因此未采用 B/S 架构进行系统设计与实现。

5.1 数据预处理界面

在数据预处理模块界面实现过程中，主要有 3 个关键点：1)文档格式的转换需对输入数据文件进行相应格式判别，而后再选择对应程序进行格式转换；2)文档数据清洗考虑情况要尽量周全，需用到正则表达式相关技术进行实现；3)对文档进行拆分过程中，需先获取列表控件中的文档行数后再进行拆分操作。其中，文档数据清洗部分是具体实现过程中的一个难点。数据预处理模块界面如图 9 所示。

图 9 数据预处理模块界面

Fig. 9 Data preprocessing module interface

5.2 标注体系界面

在标注体系模块界面实现过程中，主要有两个关键点：1)模板配置文件的读取，读取之后需要在表格控件中对实体类别、英文、英文缩写、字体颜色、背景颜色等具体内容进行设置；2)配置文件的写入，需按照自动标注部分能够识别的配置文件格式进行保存。其中，界面实现中的表格控件的数据获取及更新操作是具体实现过程中的一个难点。标注定义模块界面如图 10 所示。

图 10 标注定义模块界面

Fig. 10 Annotation system module interface

5.3 自动分析界面

在自动分析模块界面实现过程中，主要有 3 个关键点：1)领域词典的文件读取，需将词典中的实体及对应类别加载到程序中；2)自动标注的实现，需按照标准标注格式进行输出；3)标注文件中的重复性标注需进行相应去重。其中，自动标注算法和去重算法，是在具体实现过程中遇到的一个难点。自动标注用户界面如图 11 所示。

图 11 自动标注模块界面

Fig. 11 Automatic labeling module interface

输出的标注文件中，设备软件类实体在 Brat 软件中的查看效果如图 12 所示。

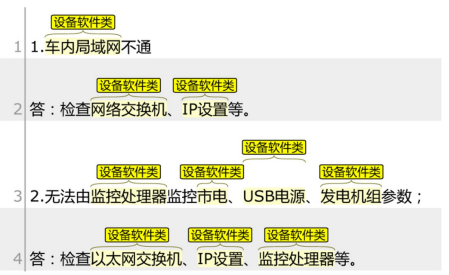

图 12 设备软件类实体标注效果

Fig. 12 Effect of equipment software entity annotation

5.4 标注分析界面

在标注分析模块界面实现过程中，主要有 3 个关键点：1)读取目录内的所有文件并显示在列表控件中；2)根据选择的具体文件，将统计分析结果的结果数据以及绘制图像显示在窗体控件中；3)对控件内的数据和图像进行更新。其中，如何根据选定的文件对表格控件和图像控件进行更新显示，是在具体实现过程中遇到的一个难点。标注分析用户界面如图 13 所示。

图 13 标注分析模块界面

Fig. 13 Annotation analysis module interface

5.5 编码转换界面

在编码转换模块界面实现过程中，主要有两个关键点：1)编码转换过程中，需根据编码选择情况进行对应的转换操作；2)样本集文件的分割数量，可以根据实际情况进行设定。其中，编码转换算法是在具体实现过程中遇到的一个难点。

编码转换模块界面如图 14 所示。

图 14 编码转换模块界面

Fig. 14 Code conversion module interface

经过 BIEO 编码转换后的命名实体识别语料的部分内容如表 4 所列。

表 4 BIEO 编码命名实体识别语料

Table 4 BIEO code named entity recognition corpus

字符序列	BIEO 编码	字符序列	BIEO 编码
—	—	—	—
车	B-EQU	无	B-FAU
载	I-EQU	法	I-FAU
计	I-EQU	启	I-FAU
算	I-EQU	动	I-FAU
机	E-EQU	、	I-FAU
操	B-EQU	报	I-FAU
作	I-EQU	错	E-FAU
系	I-EQU	。	O
统	E-EQU	\n	O
—	—	—	—

结束语 针对军事指控保障领域知识图谱构建中命名实体识别模型缺乏相应基础训练语料库的现状，本文在分析相关研究现状的基础上，设计并实现了一个基于 PyQt5 应用程序基本框架的 GUI 命名实体识别语料库构建系统。首先，简要说明了系统整体架构和语料处理技术流程；其次，详细介绍了系统的数据预处理、标注体系、自动标注、标注分析和编码转换五大功能模块的相关设计内容；最后，通过 PyQt5 应用程序基本框架和各类功能组件对各功能模块进行了图形用户界面的实现。本系统的设计与实现过程中，自动标注模型算法的实现是核心和重难点。

在未来研究过程中，应考虑结合最新的深度学习技术进一步优化自动标注模型算法，以提高自动标注的准确率。

参 考 文 献

[1] HE J Z. The concepts, reference model of C2 and its value chain analysis[J]. Fire Control & Command Control, 2019, 44(6): 1-8.

[2] HANG T T, FENG J, LU J M. Knowledge Graph Construction Techniques: Taxonomy, Survey and Future Directions[J]. Computer Science, 2021, 48(2): 175-189.

[3] LI Y, HE Y Q, QIAN L H, et al. Chinese Nested Named Entity Recognition Corpus Construction[J]. Journal of Chinese Information Processing, 2018, 32(8): 19-26.

[4] YANG J F, GUAN Y, HE B, et al. Corpus construction for named entities and entity relations on Chinese electronic medical records[J]. Journal of Software, 2016, 27(11): 2725-2746.

[5] ZAN H Y, LIU T, NIU C Y, et al. Construction and Application of Named Entity and Entity Relations Corpus for Pediatric Diseases[J]. Journal of Chinese Information Processing, 2020, 34(5): 19-26.

[6] MO T J, LI R, YANG J X, et al. Construction of named entity corpus for highway bridge inspection domain[J]. Journal of Computer Applications, 2020, 40(S1): 103-108.

[7] ZHOU B B, ZHANG H J, ZHANG R, et al. Construction of Military Corpus for Entity Annotation[J]. Computer Science, 2019, 46(S1): 540-546.

[8] FENG L L, LI J H, LI P F, et al. Constructing a Technology and Terminology Corpus Oriented National Defense Science[J]. Journal of Chinese Information Processing, 2020, 34(8): 41-50.

[9] ZHANG K. Research on semi-automatic tagging of Geographical Entities Information based on incremental learning[D]. Nanjing: Nanjing Normal University, 2020.

[10] STENETORP P, PYYSALO S, TOPI'C G, et al. BRAT: a Web-based Tool for NLP-Assisted Text Annotation[C]// Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. 2012.

DU Xiao-ming, born in 1970, Ph.D, professor, Ph.D supervisor. His main research interests include NLP and knowledge graph.

YUAN Qing-bo, born in 1989, postgraduate. His main research interests include NLP and knowledge graph.