doi:10.3772/j.issn.2095-915x.2025.04.002

面向实体识别的航空情报领域语料库构建方法

开放科学（资源服务）标识码（OSID）

赖欣　张恒嫣　唐凯　梁译丹

中国民用航空飞行学院 成都 641400

摘要：[目的 / 意义] 航空情报服务是航空运行信息的重要来源，覆盖大量相关领域知识与专业词汇，当前航空情报领域尚无公开语料库，面向实体识别展开语料库构建技术的研究有助于航空情报的数字化转型。[方法 / 过程] 以航空情报汇编资料入手，参照 AIXM 对航空实体及属性关系的定义，在领域本体框架构建的基础上建立航空情报领域实体标注体系、制定标注规范，并展开批量实体标注与一致性检验工作，形成面向实体识别任务的航空情报领域语料库。针对实体识别任务，采用 BiLSTM-CRF 模型融入领域简缩字词典，对自建语料库质量进行检验。[局限] 当前研究多采用机场类航空情报数据，一定程度上限制了实体识别模型的进一步优化和语料库规模的扩展。[结果 / 结论] 融合简缩字词典后的模型准确率、召回率、F 值分别提高了 4.18%、2.70%、3.42%，实验结果表明，该方法能够提升航空情报语料库中的命名实体的识别效率，为后续航空情报信息的实体识别和大规模语料库构建提供支撑。

关键词：语料库构建；航空情报；实体标注体系；命名实体识别

中图分类号：U8；V219；G35

A Corpus Construction Method for Entity Recognition in the Field of Aeronautical Information

LAI Xin　ZHANG Hengyan　TANG Kai　LIANG Yidan

Civil Aviation Flight University of China, Chengdu 641400, China

Abstract: [Objective/Significance] Aeronautical Information Service is an important source of aviation operation information, covering a large amount of knowledge and professional vocabulary in related fields. Currently, there is no public corpus in the field of aeronautical information, and research on corpus construction technology oriented to entity recognition is conducive to the digital transformation of aeronautical information. [Methods/Processes] Based on the data of aeronautical information, this

基金项目　中国民用航空飞行学院研究生科研创新基金项目“航空情报领域语料库构建关键技术研究”（24CAFUC10190）；中央高校校级重点项目“大数据环境下航空情报信息特征提取与事件关联研究”（ZJ2023-003）；四川省自然科学基金项目“基于异构低空网格的四维航迹规划关键技术”（2023NSFSC0903）。

作者简介　赖欣（1977-），博士，教授，主要研究方向为航空、大数据与自然语言处理，E-mail：lxrzg@163.com；张恒嫣（2000-），硕士研究生，主要研究方向为自然语言处理、航空信息管理；唐凯（2001-），硕士研究生，主要研究方向为航空信息管理；梁译丹（2002-），硕士研究生，主要研究方向为航空信息管理。

引用格式　赖欣，张恒嫣，唐凯，等 . 面向实体识别的航空情报领域语料库构建方法 [J]. 情报工程，2025, 11(4): 15-25.

doi:10.3772/j.issn.2095-915x.2025.04.002

paper establishes the entity labeling system and formulates the labeling specification in the field of aeronautical information by referring to the definition of Aeronautical Information Exchange Model and classification of aviation elements. On the basis of text preprocessing, bulk entity annotation and consistency checking were carried out to form the Aeronautical Information Service domain Corpus for entity recognition tasks. Aiming at the entity recognition task, BiLSTM-CRF model was integrated into the domain abbreviated word dictionary to test the quality of the self-built corpus. [Limitations] Current studies mostly use aerodrome aeronautical information data, which to some extent limits the further improvement of entity recognition model and the expansion of corpus scale. [Results/Conclusions] The model accuracy rate, recall rate and F-value after the integration of the abbreviated word dictionary were increased by 4.18%, 2.70% and 3.42% respectively. The experimental results show that this method can improve the recognition effect of named entities in aeronautical information corpus, and provide support for subsequent entity recognition of Aeronautical Information Service and large-scale corpus construction.

Keywords: Corpus Construction; Aeronautical Information; Entity Labeling System; Named Entity Recognition

引言

航空情报服务是保障航空运行的重要支撑服务，为民用航空运输提供航空数据和运行决策信息。当前全球航空情报运行模式正由面向产品的传统的航空情报服务（Aeronautical Information Service，AIS）向面向数据管理的航空情报管理（Aeronautical Information Management，AIM）模式过渡[1]。AIS 模型下传统的航空情报工作以人工判读信息为主，随着全球航空运行的快速发展，大量影响航空规划与每日运行的信息亟待进行数字化转化，而高效数字化转化的基础是领域语料库的建设。

语料库是指为一个或者多个应用目标而专门收集的，有一定结构的、有代表的、可被计算机程序检索的、具有一定规模的语料集合[2]。近年来，随着大数据与深度学习的发展，语料库的构建与研究逐渐成为热点，语料库构建技术进展快速，绝大多数自然语言处理（Natural Language Processing，NLP）任务如信息抽取[3]、机器翻译[4]、智能决策[5]等都是在语料库基础上训练完成的。语料库应用领域广泛，在国防军事领域，冯鸾鸾等[6]提出基于字词单元和语言学特征的技术和术语识别方法，通过人工标注方式构建面向国防科技领域的技术和术语语料库，单义栋等[7]根据军事命名实体特点提出一种基于多标签的标注方法，提升了军事命名实体的识别效果；在医疗领域，叶娅娟等[8]从糖尿病电子病历入手，采用 BiLSTM-CRF 和 RoBERTa 模型构建实体及关系标注语料库。我国民航领域语料库建设集中在对陆空通话的研究[9]，针对航空领域语料库构建技术的研究还处于起步阶段，语料规模及相关技术有很大发展空间。

航空情报语料库的构建与实体识别技术可应用于航空情报自动化处理、飞行安全监控等场景。例如，通过识别机场设施、通信设备等实体，系统可自动生成飞行通告或实时监控航行数据的合规性。由于航空情报直接影响飞行安全，其词库须符合严格的准确性与一致性要求。因此，研究如何从海量航空信息中准确抽取领域相关技术及术语，有助于加快行业数字化进程，服务于我国智慧民航建设。本文将针对航空情报产品信息特征及其特殊性，参考国

际统一标准的航空情报交换模型制定相应标注体系与标注规范，并基于深度学习模型融合领域词汇特征，最终构建面向航空情报领域的语料库。

1 语料库总体框架

1.1 语料库结构设计

航空情报领域语料库（Aeronautical Information Service domain Corpus，AISC）的构建主要包括语料收集、语料预处理、语料的序列标注以及语料库的实现 4 个部分，本文所构建的语料库总体结构设计如图 1 所示，首先对所搜集航空情报原始语料进行解析，进行文本规范化等预处理工作；基于航空情报信息交换模型（Aeronautical Information Exchange Model，AIXM）的规范构建标注体系并形成适用于本研究的标注准则，采用 BIO 序列标注得到大量标注语料；最后融合词汇特征进行实体识别实验，检验自建语料库质量；在整个语料库构建过程中，最终可以得到领域本体库、情报简缩字词库、平行双语语料库等不同存储形式与内容的领域语料库，共同服务于航空情报语料资源建设，从而推动航空情报产品这一特定领域的信息结构化、数字化进程。

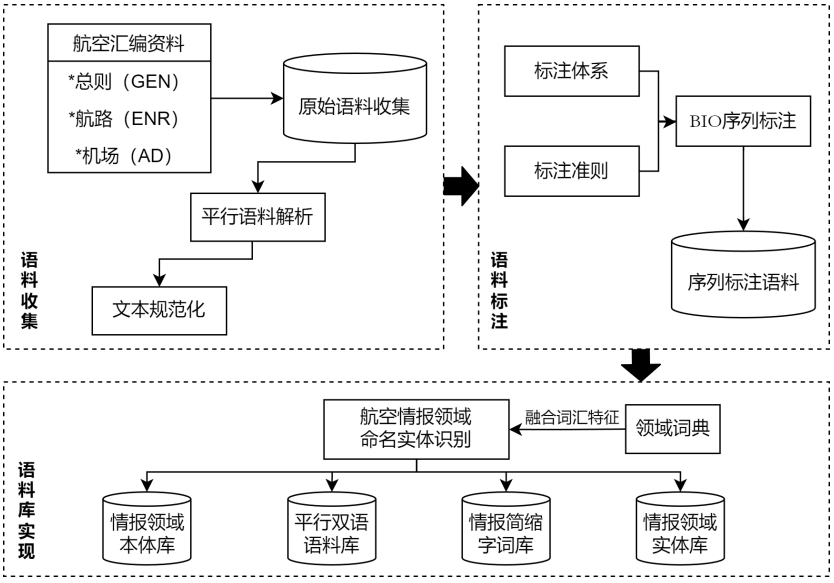

图 1 航空情报领域语料库总体结构

1.2 语料收集

作为保障航空运行安全所必须的资料，航空情报产品资料的解析与数据挖掘成为当前亟待解决的问题，本文将以航空情报产品中信息覆盖最全面的航空资料汇编（Aeronautical Information Publication，AIP）作为主要数据来源展开研究。航空资料汇编是根据国际航空公约，由国家发行或授权发行，是保障航空器在该国安全运行所必需的持久性航行资料，由总则（General，GEN）、航路（En-Route，ENR）、机场（Aeronautical，AD）3 部分内容构成，以表格、文本和图像形式发布，数据结构多样。语料的内容主要为网络公开资源以及航空公司等行业内单位提供的运行数据，具有

doi:10.3772/j.issn.2095-915x.2025.04.002

一定权威性与代表性。

1.3 语料预处理

所获取的原始语料资源多为半结构化与非结构化数据，数据结构不统一，存在双语数据等情况，对信息的抽取与结构化处理存在较大难度，因此，为保障标注工作与实体识别工作的质量，首先需要对语料进行文本解析与规范化处理。

语料库按其语种进行划分可以分为单语语料库、平行语料库、多语种语料库等[10]，其中平行语料库包含两个单语语料库且互为另一语料库的翻译，要求对相应段落或句子进行语料对齐。航空情报服务发布国内外航空运行信息，数据的共享与交换是其作为全球性情报产品的重要特征，因此，研究航空情报资料的中英双语句对齐对于领域语料库的迁移应用、平行语料库构建、智能机器翻译等自然语言处理任务具有重要意义[11]，同时，对航空情报语料进行分句处理有助于后续评估句长特征对于实体识别模型的影响。

本文所选取的语料资源主要来源于航空情报单位所发布的航空情报产品与运行信息，数据来源广泛，且多为双栏格式存储的 PDF 文件，相较于 txt 文本的解析较为困难，同时需要去除页面内头尾干扰因素及表格信息，保障语料质量与一致性，提取本文所需中文语料。因此，本文利用 Fitz 库设计了一种基于坐标快定位的文本提取与规范化算法，具体步骤如下：

第一步：遍历所有 PDF 原始语料中的每一页，获取页面矩形区域；

第二步：根据页面顶部和底部信息位置及中线位置，进行坐标快定位；

第三步：裁剪去除顶部和底部干扰信息，左右输出的文本分别为中文语料（CH_Corpus）与英文语料（EN_Corpus）；

第四步：利用正则表达式（Regex）得到去除表格信息后的所需的文本语料资源；

经过上述算法所得语料文本，基本能够满足航空情报领域实体语料库构建的需要。所得航空情报平行双语语料处理结果如表 1 所示。

表 1 航空情报平行语料库抽取结果

CH_Corpus	EN_Corpus
1.1 禁止未安装二次雷达应答机的航空器起降；	1.1 Take off/landing of aircraft without SSR transponder are forbidden;
1.2 所有技术试飞需事先申请，并在得到空中交通管制部门批准后方可进行；	1.2 Each and every technical test flight shall be filed in advance and conducted only after clearance has been obtained from ATC;
1.3 可使用最大机型：A380 及其同类机型；	1.3 Maximum aircraft to be available: A380 and equivalent;

2 领域语料库构建关键技术

2.1 标注体系构建

语料库构建的核心工作是制定完善的标注体系与标准规则[12]。在通用语料库中，实体类别通常包含人名、地名、组织机构名等，广泛适用于多种命名实体识别（Named Entity Recognition，NER）任务。但在民航领域，通用实体标签难以覆盖大量专业术语，因此，本文参照全球标准化的 AIXM，其包含航空情报信息概念模型（Aeronautical Information Conceptual Model，AICM）与交换模型两部分，

AICM 从航空信息表达的属性定义、关系定义等方面进行了规范化表示，将航空信息分为机场、空域、点、航路、程序、服务 6 个部分，并对各类要素进行明确的层级分类和定义，是得到国际认可的航空实体表达方式 [13]。为确保语料库建设的一致性和规范性，本文参照 AIXM 对领域知识的规范化描述，以机场类航空情报信息为例，利用资源描述框架（Resource Descriptive Framework，RDF）理论对航空情报语料进行规范化的语义表示，为领域知识抽取与知识图谱构建提供理论基础。图 2 为基于 Protégé 的机场类航空领域本体框架模型。

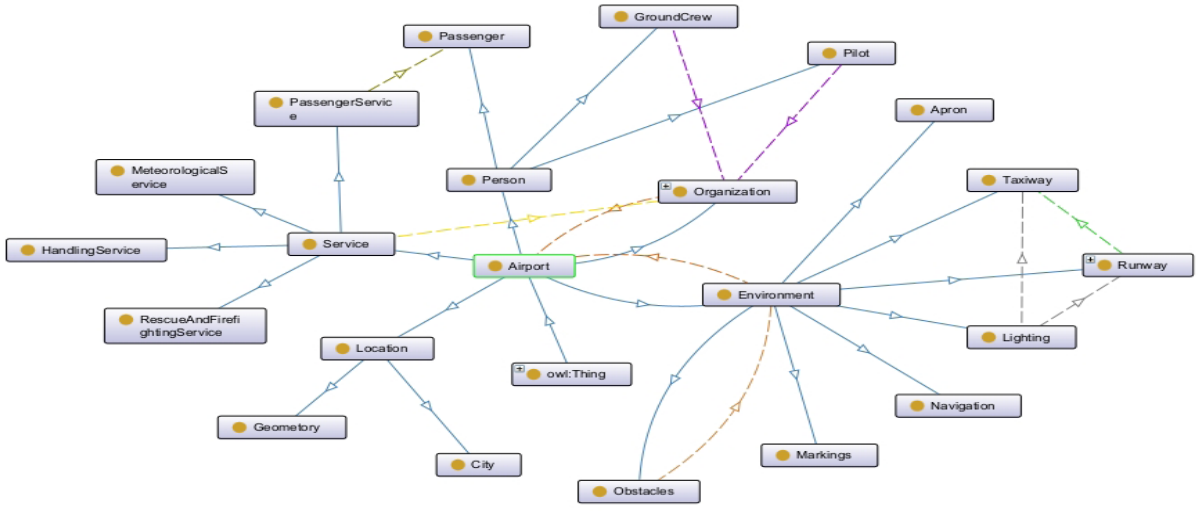

图 2　机场类航空情报领域本体层次架构

本文结合 AICM 中对实体属性概念的定义，结合领域专家意见，构建围绕航空情报语料统一的标注体系与标注准则，对领域术语实体进行层级结构划分，总结得到 6 类命名实体。6 类命名实体同时包含通用类领域实体，兼具领域适用性与行业可迁移性，下文将对本文所构建的标注体系进行详细阐述：

（1）人员（标注标签 PER），包括机场地面、空勤、空中交通管制等涉及机场运行的各类人员，同时包括人员相关资质、证书，如“空中交通管制员”“航空器驾驶员”等。

（2）组织（标注标签 ORG），包括各级省市地区管理部门、机场及其下属单位等，包括划设的各级情报区、管制区域，如“空中交通管制部门”“大兴机坪”等。

（3）活动（标注标签 ACT），指航空器在机场地面及起降阶段在机场管制范围内的各项活动、针对保障航空器运行的活动、各项空中交通服务、服务所提供的情报信息等，如“航空器进近”“慢车除冰”“告警服务”等。

（4）机场道面与灯光系统（标注标签 APT），本研究中简称机场，泛指机场飞行区、地面保障区、道面等区域的设施设备和灯光系统，如“停机位”“滑行道”“跑道中线灯”等。

（5）通信导航监视（标注标签 COM），本文中简称通导，包括机场 / 直升机场内保障

doi:10.3772/j.issn.2095-915x.2025.04.002

航空器运行的通信、导航与监视设施设备，如“二次雷达”“甚高频通信设备”等。

（6）航空器及机载设施（标注标签AIR），本文中简称航空器，指构成航空器的零部件及机载设施设备，也包含航空器飞行性能相关参数，如“航空器翼展限制”“重型机”“机载 HUD”等。

采用三标记 {B, I, O} 法进行序列标注，B 表示实体开头，I 表示实体内部，O 表示非实体词汇的组成部分，相对应的标注标签如表 2 所示。

表 2　航空情报领域语料库序列标注体系

Entity type	Entity head	Entity interior and tail
人员	PER-B	PER-I
组织	ORG-B	ORG-I
活动	ACT-B	ACT-I
机场	APT-B	APT-I
通导	COM-B	COM-I
航空器	AIR-B	AIR-I
非实体	O	O

2.2　标注准则及标注过程

由于自然语言本身具有主观性与歧义性，很多情况下针对语义的判断与分类标准具有争议，标注结果取决于标注者对领域知识的掌握与理解，而标注的一致性与准确性直接影响了语料库的质量，因此在正式展开标注工作之前，通常需要制定一套标注准则以规范争议性语料文本。命名实体标注基本准则即：（1）不重叠标注；（2）不嵌套标注；（3）实体标注尽量不包含标点及连接词。

本文结合语料信息特征制定了具体的标注准则如下：

（1）遵循命名实体标注准则。不重叠标注即同一实体不可同时具有多个标签；不嵌套标注即特殊情况如“地 / 空无线电话音通信”等实体内必要标点允许标注。

（2）遵循最大标注范围原则。如“民用固定翼涡轮发动机”中“民用固定翼涡轮”是针对“发动机”的限定词，当设施设备或机场活动类语料包含限定词或修饰语时，将其作为整体进行标注。

（3）单个高频语料不标注。如“航空器”“跑道”“滑行道”等词频数高（参考词频统计结果）且独立出现时，不进行标注，仅标注如“进港航空器”“脱离跑道”等复合词。

在建立标注体系与标注准则基础上，实施两阶段标注法，即将标注过程分为试标注阶段与正式标注阶段，对不一致的部分进行商议后修正，并进行标注一致性的评估，以提高语料标注质量。标注工作的具体流程如图 3 所示。

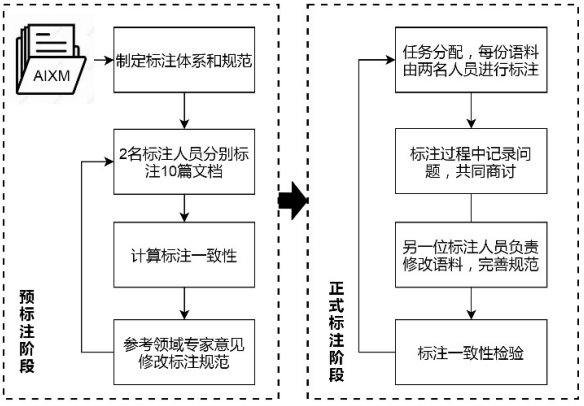

图 3　两阶段标注过程

2.3　简缩字实体词典构建

民航专业用语有其特殊性，常在各类信息

表达中采用专业技术术语或民航简缩字进行表达，各国常用航空简缩字都在GEN2.2进行呈现。例如，“出港航班机组申请ATC放行许可应不早于预计起飞时间（ETD）前20min”，句中“ATC放行许可”与“预计起飞时间（ETD）”均为本文所定义的命名实体，通过对语料库中6219条文本进行统计，共发现包含简缩字的样本1613条，约占总数26%，语料库中大量简缩字及其组合词会造成非领域实体的误判，导致准确率和召回率偏低。

对于此类简缩字的抽取问题，当前主要采用基于词库对比和基于规则统计两种方法[14]，通过建立简缩字对或内部规则完成简缩字到原实体之间的映射，实现包含简缩字实体的抽取任务。为解决上述问题，本文采用基于词库与规则统计相结合的方法，通过对航空情报产品中使用的简缩字（GEN2.2）进行提取与过滤，将得到的1145个简缩字对按特征类型划分为3类，如表3所示。

表3　情报产品简缩字类型

类型	说明	样例
独立型	独立出现，语义表示明确且唯一	ADS-B，广播式自动相关监视
复合型	与数字或字母等组合得到的实体	D…，危险区（后接识别符号）
多义型	同一简缩字表达多种语义	ARR，到达/到达电报

其中独立型简缩字能够单独作为实体进行词库匹配；复合型简缩字对需要与数字或字母进行组合，因此采用正则表达式建立组合规则；多义型则需要根据上下文语义特征判断词性进而确定词义。将3种类型简缩字组合规则及其标注序列作为词汇特征融入语料库，能够有效提高模型的实体识别效率。

2.4　语料库信息统计

2.4.1　标注数量统计

经过迭代标注与检查修正，对标注好的语料数据进行整理并统计，共提取6219条航空情报双语句对，所得10045个标注实体数目的分布情况如图4所示。

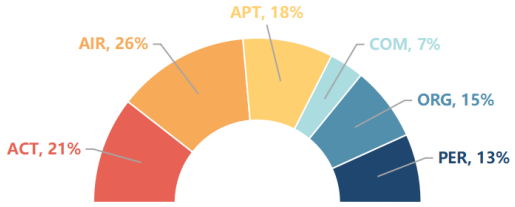

图4　AISC实体数目统计

AISC语料库覆盖航空汇编资料总则、航路和机场3部分内容，从实体数目来看，航空器及其机载设施（AIR）与活动（ACT）两个类别占比最多，分别为26%和21%，覆盖了近半数的实体，且样本涉及机场细则文本量较多，因此实体类别多集中在机场道面、活动等范围内，涵盖人员（PER）与通信导航监视设施（COM）的实体数量较少，分别仅占总数的13%与7%。

从图中的整体分布来看，实体类别的数目相对均衡，表明本文所构建的标注体系能够基本覆盖航空情报产品资料的各类航空要素及术语，且能够延伸至其他航空信息，展开命名实体识别任务，扩大语料库规模。

2.4.2　标注质量统计

语料库的标注质量通常选用Kappa值[15]和F_1值[16]进行一致性检验，而Kappa值一般适用于正例与负例比较均衡的场景，由于实体标注语料中存在大量未标注文本即负例，因此本文

doi:10.3772/j.issn.2095-915x.2025.04.002

选用 F_1 值评价多人在进行人工标注工作时的一致性。经过多轮迭代与规范的完善后，选择 50 份标注语料，将规范制定者（A）的标注结果视为标准语料，将标注人员 B 的标注结果与其进行比较，计算结果的准确率（P）、召回率（R）、F_1 值。经过统计，本文构建的航空情报领域语料库的标注一致性达到了 89.12，表明本文自建语料库的质量是可信赖的，能够为下游自然语言处理任务提供支撑。

3 命名实体识别实验

为分析自建语料库质量以及对模型性能的影响，本章利用自然语言处理技术，提出一种融合词汇特征的航空情报领域实体识别模型，对航空情报产品信息进行实体识别实验验证。

3.1 模型结构

在中文序列标注任务中，通常需要考虑上下文信息，而传统基于统计的机器学习模型如 CRF[16]、HMM[17] 仅考虑预测文本与标签之间的关系，因此本文选用 BiLSTM-CRF 融合词汇特征作为航空情报语料的实体识别模型（BiLSTM-CRF+Abb），该模型包含：Embedding 层、BiLSTM 隐藏层和 CRF 层，BiLSTM 层可以自动学习语料的上下文特征，引入 CRF 层来学习序列标注的约束条件，通过转移特征考虑输出标签之间的序列来保证预测结果的有效性，并利用外部词典，在不改变模型架构的情况下通过词典提升领域实体识别任务的性能，模型的结构示例如图 5 所示。

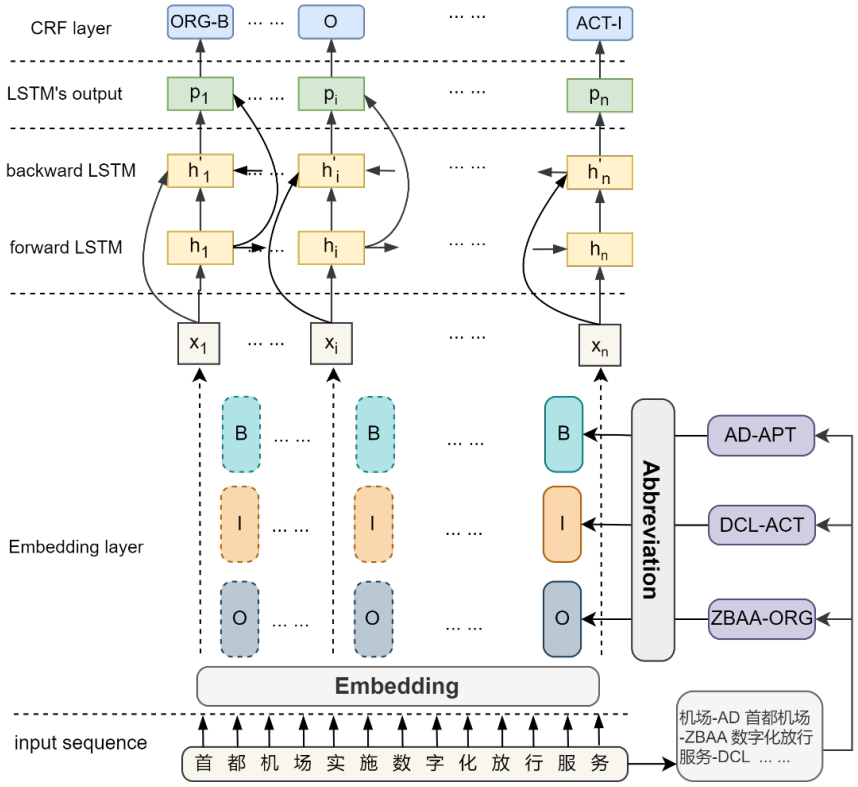

图 5 Abb+BiLSTM-CRF 模型结构

Embedding 层作为模型的输入层，通过词向量以一条航空情报语料“首都机场实施数字化放行服务”经分词后的结果为 [“首都机场”，“实施”，“数字化放行服务”] 为例，经嵌入层将句中每个单词映射到一个嵌入空间，转化为随机初始化向量表示：

$$X=(X_1,X_2\cdots X_n),x_i\in R^d \quad (1)$$

其中 n 为输入序列长度，经过输入层将其转化为词向量，向量矩阵 $n\times d$ 中 d 为模型参数中的 Embedding 层维数。

同时，在 Embedding 层，基于 GEN 2.2 标准建立 3 类正则模板，通过并行标注通道将规则匹配结果转化为 BIO 格式的辅助标签序列，进行序列标注层的叠加：

①独立型模板：采用精确匹配策略；

②复合型模板：设计结构化规则；

③多义型模板：结合上下文约束。

这种方法将词汇特征信息纳入字符表示，只需调整字符表示层引入词典信息，不改变模型原本的架构。

隐藏层采用的双向长短时记忆网络包含前向、后向两个 LSTM 网络，将输入层的航空情报语料转化为向量形式，将向量序列 X = (x_1, x_2,⋯, x_n）输入 Bi-LSTM 中，相关公式表示为：

$$h_i=f(UX_i+Vh_{i-1}) \quad (2)$$

$$h_i^{'}=f(U^{'}X_i+V^{'}h_{i-1}^{'}) \quad (3)$$

$$p_i=g(Wh_i+Wh_i^{'}) \quad (4)$$

其中，p_i 表示输出序列，g 表示激活函数，作用于隐藏层到输出层，h_i 和 h'_i 分别表示前向层和后向层的输出，f 表示激活函数，作用于隐藏层内部，W、U、V、U' 和 V' 均表示权重，X_i 表示输入序列的第 i 个字符，h_{i-1} 表示 i–1 时刻前向 LSTM 的输出，$h_{i-1}^{'}$ 表示 i–1 时刻后向层的输出，同时输出序列 p_i 作为 CRF 层的输入，对标签序列进行解码，将输出序列 p_i 的维度映射到 K 维（表示标签类别数量，本任务中 K=6），并通过 softmax 函数进行归一化计算，计算公式为：

$$\mathrm{Y}_i=\mathrm{softmax}(\omega H_i+b) \quad (5)$$

其中 Y_i 表示模型对第 i 个输入单元标签预测的概率分布，ω 表示权重，b 为偏置项。

最终的输出结果为公式（5）中 Y_i 的最大值对应的标签，为保证整个输出序列结果的有效性，引入 CRF 层用于建立标签的规则约束，例如 ORG-B 的标签后一定接 ORG-I 而非 PER-I，输出序列对应预测分数如公式（6）、公式（7）所示，得到的路径分数 S(X,Y) 即为最佳输出序列标签，本文采用维特比算法进行解码。

$$\mathrm{S}(X,Y)=\sum_i^n(A_{y_i+y_{i+1}}+P_{i,y_i}) \quad (6)$$

$$Y^*=\mathrm{argmax}\,S(X,Y) \quad (7)$$

3.2 评价指标

命名实体识别任务的评价标准通常分为两部分，首先需要判断实体的边界，其次判断实体所属标签类型。本文采用准确率 P、召回率 R 和 F_1 值作为实体识别任务的评价指标，选用精确实体匹配方式，即要求实体边界及实体类别均识别正确，相关评价指标表示如下：

$$P=\frac{TP}{TP+FP}=\frac{\text{实体类别被正确识别的样本数}}{\text{识别为实体的样本数}} \quad (8)$$

$$R=\frac{TP}{TP+FN}=\frac{\text{实体类别被正确识别的样本数}}{\text{标记为实体的样本总数}} \quad (9)$$

doi:10.3772/j.issn.2095-915x.2025.04.002

$$F_1 = \frac{2 \bullet P \bullet R}{P + R} \quad (10)$$

3.3 实验结果及分析

由于标注语料规模较小，为评估自建语料库 AISC 在命名实体识别模型中的效果，本文采用双向长短时记忆网络 + 条件随机场（BiLSTM-CRF）模型，同时加入融合简缩字词典特征的 BiLSTM-CRF 模型（BiLSTM-CRF+Abb）作为对比，即 2.3 节构建的简缩字词典及其分类标签，通过调整模型 Embedding 层，将规则匹配模板转化为 BIO 辅助标签序列，进行标注序列层的叠加，进行预训练领域词汇增强嵌入，提高模型在特定领域任务中的表现。

对相同数据集进行命名实体识别，将 6219 条经过去重的航空情报语料数据按照 7:2:1 比例划分训练集、验证集和测试集，模型的参数设置如表 4 所示。

表 4　实验参数设置

参数	训练值
Batch_size	64
Learning_rate	0.001
Epoch	32
Emb_size	128
Hidden_size	128

表 5 给出两个模型在各类实体上的识别效果与总体情况，可以看出通用领域实体人员与组织机构的识别效果最好，是由于这两类实体边界与构词方式相对明确，如人员类别多以“XX 员”“XX 人”“XX 机组”形式出现，组织机构类多以“XX 部门”“XX 中心”“XX 管制区”为后缀，而活动、机场两个类别由于覆盖实体面较广，在标注过程中容易漏标，且涵盖大量领域简缩字实体，因此识别效果较差。而加入领域简缩字特征后，相较于 BiLSTM-CRF 模型，准确率、召回率、F_1 值分别提高了 4.18%、2.70%、3.42%，表明该方法一定程度解决了在标注过程中难以对简缩字人工进行逐一检索，以及模型对于简缩字实体无法进行上下文的语义识别、对实体边界的识别存在误差等问题。

表 5　命名实体识别的对比结果

实体类别	BiLSTM-CRF			BiLSTM-CRF+Abb		
	P/%	R/%	F_1/%	P/%	R/%	F_1/%
APT	85.8	76.9	81.1	85.8	85.9	85.8
AIR	83.1	87.8	85.4	88	86.9	87.4
ACT	78.6	84.8	81.6	86.1	85.1	85.6
COM	85.4	83.5	84.4	90.4	82.4	86.2
ORG	87	79.4	83	93.1	**89.3**	**91.2**
PER	**92.5**	**88.3**	**90.4**	**94.1**	87.3	90.6
总体	85.4	83.5	84.3	89.6	86.2	87.8

4　总结与展望

本文围绕航空情报文本的结构化处理与语料库构建，构建了完整的技术体系，主要内容及创新成果如下：

（1）在对航空情报产品数据特点进行分析的基础上参考了 AIXM 规范，制定了标注体系与标注准则，提出一种面向实体识别的航空情报领域语料库构建方法。

（2）采用命名实体识别模型融合领域词典特征进行实验对比，检验了自建语料库的质量。

（3）首次针对航空情报领域、融合情报产品简缩字构建面向实体识别任务的领域专用语料库，为航空情报产品结构化、数字化建设奠定基础。

本文仍然存在部分局限性，本文所构建的航空情报领域语料库主要采用 AIP 文本数据，语料库覆盖范围有限，由于领域内标注数据集稀缺，本文所建立的命名实体识别模型在嵌套实体、非规范性表述等复杂任务下的处理能力不足。

因此，在后续研究中，对航空情报多源异构数据的处理、信息的实时交互将成为未来数字化航空情报发展的趋势，未来可从迁移学习角度，利用数据增强等方法提高低资源语料库的实体识别质量，为大数据驱动下的航空情报信息处理提供高质量的语料库支撑。

参 考 文 献

[1] 中国民用航空局 . 中国民航航空情报管理（AIM）实施指南 [R]. 北京：中国民用航空局空中交通管理局 , 2018.

[2] 黄水清 , 王东波 . 国内语料库研究综述 [J]. 信息资源管理学报 , 2021, 11(3): 4-17, 87.

[3] NOGUERA-DÍAZ Y, PÉREZ-PAREDES P. Register analysis and ESP pedagogy: noun-phrase modification in a corpus of English for military navy submariners[J]. English for Specific Purposes, 2019, 53(1): 118-130.

[4] CHANG D, CHEN M, LIU L, et al. Diakg: an annotated diabetes dataset for medical knowledge graph construction[C]// Proceedings of the 6th China Conference on Knowledge Graph and Semantic Computing, 2021: 308-314.

[5] LEVSHINA N. Corpus-based typology: applications, challenges and some solutions[J]. Linguistic Typology, 2022, 26(1): 129-160.

[6] 冯鸾鸾 , 李军辉 , 李培峰 , 等 . 面向国防科技领域的技术和术语语料库构建方法 [J]. 中文信息学报 , 2020, 34(8): 41-50.

[7] 单义栋 , 王衡军 , 王娜 . 基于多标签的军事领域命名实体识别 [J]. 计算机科学 , 2019, 46(S2): 9-12.

[8] 叶娅娟 , 胡斌 , 张坤丽 , 等 . 糖尿病电子病历实体及关系标注语料库构建 [J]. 中文信息学报 , 2023, 37(12): 17-25.

[9] 李荣枝 . 航行数据语料库研究与构建 [D]. 德阳：中国民用航空飞行学院 , 2022.

[10] CHEN J, NIE J Y. Automatic construction of parallel English-Chinese corpus for cross-language information retrieval[C]// Proceedings of the 6th Conference on Applied Natural Language Processing, 2000: 21-28.

[11] 温松 . 中英平行语料库关键技术研究及应用 [D]. 南宁：广西民族大学 , 2021.

[12] 周彬彬 , 张宏军 , 张睿 , 等 . 面向实体标注的军事语料库建设 [J]. 计算机科学 , 2019, 46(S1): 540-546.

[13] 胡敬玉 . 基于 AIXM 的航空要素编码规则与可视化研究 [D]. 德阳：中国民用航空飞行学院 , 2020.

[14] 崔世起 , 刘群 , 林守勋 , 等 . 中文缩略语自动抽取初探 [C]// 全国第八届计算语言学联合学术会议（JSCL-2005）, 2005: 65-70.

[15] CARLETTA J. Assessing agreement on classification tasks: the kappa statistic[J]. Computational Linguistics, 1996, 22(2): 249-254.

[16] HRIPCSAK G, ROTHSCHILD A S. Agreement, the F-measure, and reliability in information retrieval[J]. Journal of the American Medical Informatics Association, 2005, 12(3): 296-298.

[17] BIKEL D M, SCHWARTZ R, WEISCHEDEL R M. An algorithm that learns what's in a name[J]. Machine Learning, 1999(34): 211-231.