面向实体标注的军事语料库建设

周彬彬 张宏军 张 睿 冯蕴天 徐有为
(中国人民解放军陆军工程大学指挥控制工程学院 南京 210007)

摘 要 军事语料的识别和标注是军事语料库建设的关键。针对军事语料的实体,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。该框架借助设计的高精分类器进行基本特征的选择和提取,结合军语的典型特征组成特征集,构建基于军语词典校正的特征空间,对军事语料进行实体识别之后按照指定的标注规范和词形标记规范进行军事语料实体的标注,构建一个较大规模的高质量军事语料库。实验表明,该框架可以较好地完成语料实体的识别和语料标注工作,有利于军事语料库的建设工作和认清其在军事上的广泛作用和应用前景。

关键词 军事实体标注,军语词性标记,特征提取,军事语料库
中图法分类号 TP391 **文献标识码** A

Construction of Military Corpus for Entity Annotation

ZHOU Bin-bin ZHANG Hong-jun ZHANG Rui FENG Yun-tian XU You-wei
(School of Command and Control Engineering, Arm Engineering University of PLA, Nanjing 210007, China)

Abstract The key to build military corpus are the identification and the marking of military corpus. For the entities of military corpus, this paper put forward a set of unified army language part-of-speech tags specification and military corpus annotation specifications, and designed a kind of automatic extension of military corpora based on the military language dictionary entity framework feature extraction. With the help of high precision classifier, the framework selects and extracts the basic features, combined with the typical features of the language set, builds the feature space. Based on the language dictionary correction for military corpora entity recognition, according to the specified annotation standard and specification of morphological marker military annotation corpus entity, the framework builds a large-scale high-quality military corpus. Experiments show that the framework can better complete corpus entity recognition and corpus annotation of the work, to do the construction of military corpus work and to recognize its function and the application prospect of widely in the military.

Keywords Military entity's annotation, Military speech tagging, Feature extraction, Military corpus

1 概论

军事领域语料库是内容涉及军事(或为军事服务)的单一语种或多语种的文本所组成的标注语料库[1],按具体用途可分为军事通用语料库、军事专用语料库和军事服务语料库等。随着信息时代的来临,传统的战争形式已经由依靠机械化军队取胜的机械化战争转变为利用信息和信息技术、信息系统、信息化武器装备等,并通过多种手段在战场上以夺取和建立一定时空范围内战场信息优势为核心的一体化军事行动对抗的信息化战争的方式。随着信息化部队的实现和建设,大量的信息系统和数字化装备器材等已经投入到部队各层的军事行动中,随之也产生了大量的军事信息。信息化战争的核心是对信息资源的争夺与占有,信息匮乏或信息弱势的一方注定会成为战争的输家[2]。我们可以借助语料库对大量信息进行检索与分析,从中获得主要谍报,为相关部门和专家等提供决议计划帮助。

在信息化战争中,数据主导决策将是获取战场优势的关键,各军事领域、军兵种均将与作战指挥等相关的军事数据资源一起作为重点来建设。军事领域语料库的研究、建设与应用在大多数国家才刚刚兴起,但以美国为首的西方发达国家早已意想到军事语料库对信息化条件下的军事斗争所具有的潜伏代价,并已经在这方面进行了长时间的探索,积极开展了军事语料库的钻研和扶植实践活动,并把钻研成果转化到实战当中[3]。美国国防高级研究计划局(Defense Advanced Research Projects Agency, DARPA)的很多项目都是在充足语料资源的基础上借助自然语言处理技术进行的,如自动翻译、跨语言情报侦测、情报抽取和特定事件追踪与检测等。国内军事信息处理的研究刚刚兴起,很少有学者对这些数据进行标准化的标注,因而相关的标注语料资源较为匮乏。在充足标注语料资源的基础上,进行军用大数据的处理和分析,可以

周彬彬(1993-),男,硕士生,主要研究方向为军用数据与知识工程,E-mail:Zbb930707NJ@163.com(通信作者);**张宏军**(1963-),男,教授,博士生导师,主要研究方向为军事建模与仿真;**张 睿**(1977-),男,副教授,主要研究方向为军用数据与知识工程;**冯蕴天**(1990-),男,博士生,主要研究方向为军用知识与数据工程;**徐有为**(1994-),男,硕士生,主要研究方向为复杂系统分析与优化。

提高指挥员的知识发现能力和决策效率，更好地服务于战场信息控制与掌握，确保对敌的信息优势。只有建立标准的标注语料库才能从数据优势向决策优势跨越，提高决策水平和作战效能，因此本文利用可利用和可收集的各类信息，最终构建出一个规模较大、标注类型丰富的面向实体标注的军事语料库。

随着我军信息化建设的不断推进，加快军事数据资源建设已经成为当前我军信息化建设的主要内容。面向实体标注的军事语料库填充了目前军事领域相关语料资源的空白，具有重大理论意义和实际意义。它不仅可以更好地推动军事语料库的开发和研究，提高军事语言的研究效率，还可以在军事行动中依托语料库大大提升我军信息化作战的战斗力和生存力，在多国联合演习、国际维和和军事谈判与交流等方面有着巨大的应用前景和研究价值。在信息检索、战场情报获取和信息过滤等方面，面向实体标注的军事语料库建设也有着极大的研究需求和应用前景。面向实体标注的军事语料库的建设，可以提高对战场态势的综合信息捕捉和利用的能力，是保障信息化战争中信息优势的有力依靠和手段。

2　语料库的结构设计

面向实体标注的军事语料库的构建主要包括语料收集、预处理、语料标注、语料生成和数据应用服务 5 个部分。本语料库的总体结构设计框架如图 1 所示。其中预处理对生语料或熟语料进行文本规范化处理，以及中文分词和词性标注处理，得到存放在磁盘中的库文件；根据制定的标注规范自动扩展迭代的语料实体标注，并存入实体库中；再根据语料生成规范生出标准的 XML 标注语料，存入标注语料库中；语料库的数据应用部分则对库文件中的语料信息和军事实体信息运行特定的工具进行分析处理，实现特定的功能。

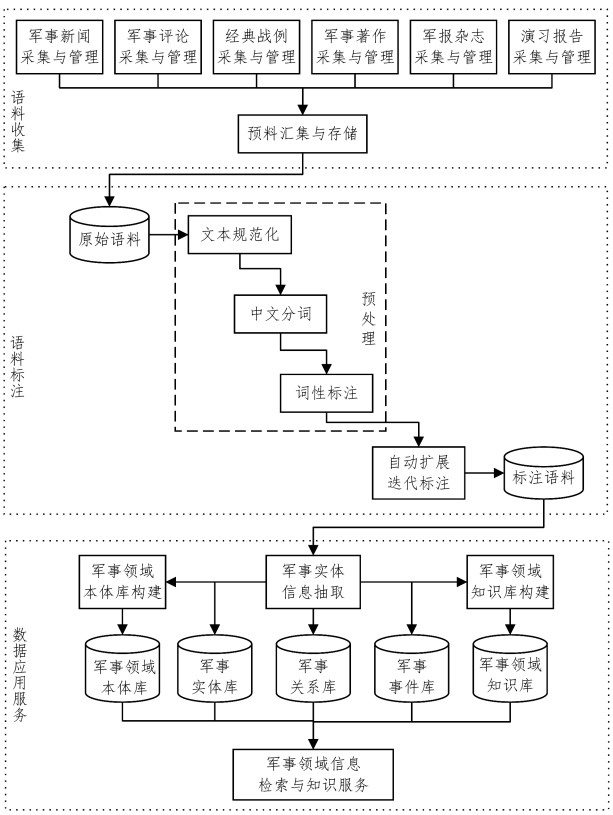

图 1　语料库总体结构设计框架

2.1　语料收集

随着信息技术的发展和公众对部队发展和建设的关注，诸多军事信息也以文本、图片、视频等方式，通过电视台、公众号、微博等手段对外展现，如军事著作、军事新闻、演习和战例分析等。现代计算机技术和网络资源使得这类语料的获得变得方便容易，但是仅仅依靠网络资源收集到的中文军事语料是不够的，我们还需要更多具有军事特性和军事特征的军事文本，如军用文书[4]。选择语料时需要对良莠不齐的文本数据进行分辨，选取高质量的军事文本作为标注对象。为了保证语料库的质量与平衡，我们从军事新闻、军事评论、经典战例、军事著作、军报杂志和演习报告 6 个来源中收集了 1040 篇军事文本，丰富了军事文本的类型。这些军事文本由于涉密导致语料难以获取，或者收取到的军事文本缺乏统一的规范，这里我们先收集这些军事文本，在之后的规范化处理中，再将其按照统一的规范进行处理。因此，获取军事语料的方式是多种多样的且需要其他部门配合收集。

2.2　预处理

2.2.1　文本规范化

我们将收集到的军事语料统一进行编码转化处理，为方便后期加工和标注，我们采用 UTF-8 编码格式，这是一种针对 Unicode 的可变长度字符编码，又称万国码，可显示各种语言。对于一些从微博和公众号上获取的军事语料，我们要对其进行处理，这是因为网页和公众号发布内容页面会存在一些无关字符、表情符号等，另外还会存在一些干扰词，如“上一页”“下一页”“赞”等。这些符号在语料中的位置一般是固定的，或者是连续的字符，因此我们收集完语料之后，要将这些干扰字符和干扰词去除。此外，军用文书类的军事语料格式信息简明清楚，因此可以直接提取这类语料的实体进行标注。

一篇语料分为篇头和篇体两部分，篇头主要包含语料的元信息，而篇体包含语料本身及其标注信息。原始语料的标签及含义如表 1 所列。其中，〈DOC〉和〈/DOC〉之间为整篇语料的所有内容；〈DOCID〉和〈/DOCID〉之间为篇头，包含语料的编码；〈BODY〉和〈/BODY〉之间为篇体，即语料的主体部分；〈HEADLINE〉和〈/HEADLINE〉之间为语料的标题；〈TEXT〉和〈/TEXT〉之间为语料的正文。

表 1　原始语料的标签及含义

标签	含义
〈DOC〉	整篇语料
〈DOCID〉	语料编码
〈BODY〉	篇体
〈HEADLINE〉	语料标题
〈TEXT〉	语料正文

其中，语料的编码包含语料的元信息。本节的编码模式仅对最必要的元信息进行采集展示，主要包括语料来源、编号和日期，为军事语料提供标识。参考了《人民日报》语料库的编码规则[5]，本节将语料编码设置为 16 位，具体为：来源分类号(4 位大写字母)＋文本编号(4 位数字)＋日期编号(8 位数字)。表 2 对一篇语料的语料编码进行了示例解析，该篇语料文件收集自军事新闻(NEWS)，其文本编号为 0001(0001)，其收集的日期为 2018 年 4 月 18 日(20180418)。

表 2 语料编码示例解析

编码	NEWS.0001.20180418
来源分类号	NEWS
文本编号	0001
日期编号	20180418

2.2.2 中文分词

预处理部分可以使用成熟的自然语言处理工具包，这些工具包虽然不是为军事领域定制的，而是面向通用领域的，但是它们可以对军事文本进行一些基础的加工处理，以辅助标注人员进行语料标注工作，也为以后进行军事实体识别和标注做好准备。对于军事语料的分词处理，我们在中科院最新研制、更新公布的 ICTCLAS2015 版本的分词系统的基本框架上，将其中的词性标记规范和标注规范剔除，引入我们设计并建立的军事词性标记规范和从军事领域专家标注的“黄金语料”中提取出的军事领域的语料实体标注规范，对收集到的生语料进行分词和词性标注处理，标注完毕的词性结果可以留作下文词性特征提取的输入。我们在进行分词处理时发现，分词的粒度标准对最终的结果有很大影响，仅靠词性特征会导致颗粒度偏大，而词形又会导致颗粒度偏小，因此在特征提取部分，需要引入一个平衡因子来进行调节，这在下文的特征提取中会有详细的描述。此外，我们在分词中加入了最大匹配分词算法，将我们在之前构建的军事词典内容进行了匹配运算，提高了分词的准确率和合理性。

2.2.3 军语词性标记

对于分词和词性标记这些简单的语言处理任务，其标记规范一般是通用的、固定的，但是其具体的划分标准、词类划分的粒度和标记符号都不统一。例如，中国科学院计算技术研究所研制的汉语词法分析器系统中采用的汉语词性标记集共计 99 个类别，其中 22 个一级标注集，66 个二级标注集，11 个三级标注集[6]。而在 LDC 标注语料中，仅一级标注集就划分了 33 个。本文在中国科学院研究制定的汉语词性标记集的基础上，将原有的通用领域的部分词性划分去除，引入具有军事特色词性标记集合，新增了 6 个一级标注集，27 个二级标注集，69 个三级标注集，将军事领域相关的命名实体的词性和类别囊括进去。

2.2.4 军事语料的标注规范和生成规范

语料构建的核心工作是制定规范和根据规范进行标注。文献[7]总结了 3 种语料标注的方法：第一种是传统的领域专家标注，标注质量高，但是成本也高，费时费力；第二种是众包标注，低成本标注大规模语料，但只能完成简单任务，如 Amazon 设计的标注平台 Mechanical Turk[8]；第三种是团体标注，从众多结果中采用信息检索评价的方式，人工解决不一致标注。

军事语料库的构建服务于军事教学和军事行动作战等，因此对其标注质量有着严格的要求，我们通过领域专家对小批量的语料进行标注，将这部分标注语料作为“黄金语料”，根据领域专家的标注制定类似“黄金语料”的标注规范，并将其制定成规范模板，加入标注平台或标注系统中，之后通过机助人工的方式，获得大规模的统一标注规范的标注语料。如此在获得高质量的标注语料的同时，又降低了语料标注的成本，提高了标注的效率。

对于军事语料库的数据文件，我们统一采用 XML 格式进行存储，并实现对军事文本实体内容的管理。如果说军事语料的标注规范是指导如何去标注军事语料中的实体内容以及该怎样标注，那么军事语料的生成规范就是去指导如何生成一份准确的、适合的 XML 标注语料文档和标注语料文档中应该包含的元素和规范，例如标注语料中元素 ID 标识符全语料库唯一，应该包含 XML 文档中必要的篇头和篇体元素等。

2.3 库文件的数据结构

将一篇原始语料中的军事实体、时间实体以及实体在军事事件中的论元角色都标注出来，单独保存，就生成了一篇标注语料。在面向实体的军事语料库建设中，每个标注元素及其属性都由文本平行标签标识出来，简洁、易懂，标注语料的标签及含义如表 3 所列。

表 3 标注语料的标签及含义

标签	含义
〈ENTITY〉	军事实体
〈EXTENT〉	具体内容
〈CHARSEQ〉	汉字序列
〈HEAD〉	实体中心词
〈RELATION〉	军事关系
〈RELATION_ARGUMENT1〉	关系的第一个变元
〈RELATION_ARGUMENT2〉	关系的第二个变元
〈EVENT〉	军事事件
〈ANCHOR〉	事件触发词
〈EVENT_ARGUMENT〉	事件论元

其中，军事实体的起始标记和结束标记分别用〈ENTITY〉和〈/ENTITY〉来表示；〈EXTENT〉和〈/EXTENT〉之间为标注元素（军事实体、军事关系或者军事事件）的具体内容；〈CHARSEQ〉和〈/CHARSEQ〉之间为汉字序列，反映标注元素的语言单位，通常是词语或句子，其子标签 START 和 END 可以用来记录标注元素的起止位置，即开头字符和结尾字符位于整个军事文本的位置，如表 3 中的第一条记录所示；〈HEAD〉和〈/HEAD〉之间为实体中心词，是军事实体的核心；军事关系的起始标记和结束标记分别用〈RELATION〉和〈/RELATION〉来表示；〈RELATION_ARGUMENT1〉与〈/RELATION_ARGUMENT1〉之间和〈RELATION_ARGUMENT2〉与〈/RELATION_ARGUMENT2〉之间分别为关系的第一个变元和第二个变元，指军事关系中涉及的两个军事实体；军事事件的起始标记和结束标记分别用〈EVENT〉和〈/EVENT〉来表示；〈ANCHOR〉和〈/ANCHOR〉之间的是事件触发词，可以指示出军事事件的类别；〈EVENT_ARGUMENT〉和〈/EVENT_ARGUMENT〉之间为事件论元，通常指涉及到的军事实体，可在一个军事事件中担任不同的角色。

子标签包含着标注元素的元信息，具体的子标签示例解析如表 4 所列。其中，〈ENTITY〉，〈RELATION_ARGUMENT1〉，〈RELATION_ARGUMENT2〉和〈EVENT_ARGUMENT〉的元素类型都是军事实体，它们的子标签 ID 为军事实体的编码，TYPE 为军事实体的大类，ROLE 为军事实体在军事事件中担任的角色；〈RELATION〉的元素类型是军事

关系，其子标签 ID 为军事关系的编码，TYPE 为军事关系的大类，SUBTYPE 为军事关系的子类；〈EVENT〉的元素类型是军事事件，其子标签 ID 为军事事件的编码，TYPE 为军事事件的大类，SUBTYPE 为军事事件的子类。上述 TYPE 和 SUBTYPE 必须严格地与本章提出的标注体系中的类型信息相对应。

表 4　子标签示例解析

标签	子标签信息	元素类型
CHARSEQ	〈CHARSEQ START="26" END="41"〉	汉字序列
ENTITY	〈ENTITY ID="NEWS. 0001. 20180418-E1" TYPE="PER"〉	军事实体
RELATION	〈RELATION ID="NEWS. 0001. 220180418-R1" TYPE="PHY" SUBTYPE="Located"〉	军事关系
RELATION_ARGUMENT1	〈RELATION_ARGUMENT1 ID="NEWS. 0001. 20180418-E4" TYPE="FAC"〉	军事实体
RELATION_ARGUMENT2	〈RELATION_ARGUMENT2 ID="NEWS. 0001. 20180418-E3" TYPE="LOC"〉	军事实体
EVENT	〈EVENT ID="NEWS. 0001. 20180418-EV1" TYPE="FLI" SUBTYPE="Attack"〉	军事事件
EVENT_ARGUMENT	〈EVENT_ARGUMENT ID="NEWS. 0001. 20180418-E1" TYPE="PER" ROLE="AGENT"〉	军事实体

本节对每个标注元素都设置了结构层次分明和表意明确的编码，即子标签 ID。一个标注元素的完整编码 ID 具体为：原始语料编码＋元素类型＋该元素在语料中的编号，其具体示例解析如表 5 所列。其中，元素类型 E 表示军事实体，元素类型 R 表示军事关系，元素类型 EV 表示军事事件；该元素在语料中的编号表示该标注元素在这篇军事文本中出现的序数。

表 5　标注元素编码示例解析

编码	原始语料编码	元素类型	该元素在语料中的编号
NEWS. 0001. 20180418-E1	NEWS. 0001. 220180418	E	1
NEWS. 0001. 20180418-R1	NEWS. 0001. 20180418	R	1
NEWS. 0001. 20180418-EV1	NEWS. 0001. 20180418	EV	1

原始语料以字为单位进行检索和统计。在依据上述编码模式生成的标注语料中，各个标注元素是相互独立的，每个标注元素都具有一个唯一的编码，以便于对标注元素进行统计、查找和修改，实现对语料内容的管理，同时也为后文训练模型前进行语料解析提供方便。制定标准的、结构化的语料编码有助于军事信息抽取语料库的机读化，具有十分重要的实际意义。

2.4　数据应用

在数据应用阶段，首先将标注的实体和标注语料存入相应的库中，作为军事语料库的基本数据库，之后根据我们识别抽取和标注的实体内容，可以建立面向军事实体的相关技术和模型研究，构建出军事领域本体库以及军事领域知识库。在这些数据库的基础上进行军事领域信息检索与知识服务，为指挥人员提供决策辅助和信息支持。最后可在语料库的基础上，开发相关的军事应用软件，以提高部队信息化作战水平和战场信息获取能力。应用服务层可以提供军事领域信息检索与知识服务等高级应用。在构建出军事实体语料库并进行军事信息分析后，可以进行军事文本可视化、军事领域信息检索、军事知识管理和舆情监控等。

3　语料库建设的关键技术

对于收集到的军事语料，都需要对其进行规范化预处理之后再进行分词处理和词形标注，军事语料实体标注主要分为特征的提取、特征集合的建立、军事语料实体的分类识别和特征的选择和结果校正。特征集合的建立是指针对军事文本和军事语料实体的特点，定义词语的基本特征并进行提取，再结合从建立的军事词表中提取的军事实体典型特征，根据我们设计的自动扩展的军事语料实体特征库模型将基本特征进行融合，从而获得完备的军事语料实体特征库。对军事语料进行军事实体的识别和标注，再根据军事语料生成规范生成标注语料，将其纳入我们建立的标注语料库中。

传统的语料库建设方法通过人工去注释文本信息，这是一项耗时的任务，需要大量的技巧和相关领域的知识。目前，基于机器学习的方法进行命名实体识别和语料标注已经成为主要方法，然而，人工注释的机器学习数据仅限于几个公共数据集，几乎完全是通用类型的新闻专线[9]。对于不同的语言和预定义的实体类型，需要相应的注释库来训练新的实体识别(NER)模型，在特定领域训练的模型往往在不可见的领域表现更差，这种数据依赖阻碍了现有的 NER 系统的适应性和军事领域可移植性。目前，国内在军事语料库建设方面还没有比较权威和统一的军事语料库，而人工收集标注的军事语料费时费力，使得建立大规模的军事标注语料难上加难。为弥补上述不足，对军事语料实体的特点进行研究，提取军事语料实体特征集，对军事语料进行语料实体标注，建立面向实体标注的军事语料库。

3.1　特征的提取

军事语料中实体的识别和标注有别于通用领域的实体，其最核心的区别就在于构建军事语料实体特征集合时，要对具有军事特色、军事涵义的实体进行特征的提取。因此，我们在建立特征集合时，除了提取由军事语料实体组建的军语词典的典型特征外，还定义了军事语料中实体的词形特征词性特征和组合特征等基本特征，并且对这些基本特征进行了提取，再将其融入扩展军事语料实体特征库中。

3.1.1　词形特征

由于我们建立了军事语料实体词典，因此词典中的词或单个字都可以单独构成一类标注实体，之前我们建立了人员军职军衔词条、军事装备词条、军用物资词条、军事设施词条、军事机构词条 5 个词典，再加上时间词、数量词、形容词、动词、量词、介词等，共计 28 个词形。我们用 W 来表示分词所得的词语组成的词序列，如式(1)所示，用 WC 表示由词形构成的序列，如式(2)所示，词形特征就是根据词形序列 WC 产生候选标注实体的。

$$W = w_1 \quad w_2 \quad \cdots \quad w_i \quad \cdots \quad w_n \tag{1}$$

$$WC = wc_1 \quad wc_2 \quad \cdots \quad wc_i \quad \cdots \quad wc_n \tag{2}$$

3.1.2 词性特征

在军事语料进行分词之后，我们可以确定每个词的词性并对其进行标注。词性特征可以根据词性序列 T 产生候选的标注实体，研究[10]表明，在建立军事语料实体特征时，将词性作为一种特征，能够大大提高其性能。根据前文介绍的词性标注规范，我们依据规范对军事语料分词所得的词语进行词性标注，通过对词性和词形特征的提取，将语料实体的识别和标注转化为一个序列化的数据标注问题，即每个句子中的词组成一个词序列，只是这个词序列带有词性标记和词形标记。我们将词性序列 T 用序列表达式的方式来表现为：

$$T=t_1 \quad t_2 \quad \cdots \quad t_i \quad \cdots \quad t_n \tag{3}$$

则带有词性标注的词序列 WT 的表达式如式(4)所示：

$$WT=w_1/t_1 \quad w_2/t_2 \quad \cdots \quad w_i/t_i \quad \cdots \quad w_n/t_n \tag{4}$$

其中，n 表示句子 S_k 被分词之后的词的个数，t_i 表示标注词的词性。将这些带有词性标注的词序列 WT 作为输入，进行语料实体的识别和标注，最后输出一个最优的“词形/词性”序列，可以用式(5)表示：

$$WC^*/TC^*=wc_1/tc_1 \quad wc_2/tc_2 \quad \cdots \quad wc_i/tc_i \quad \cdots \quad wc_m/tc_m \tag{5}$$

其中，$m\leqslant n$，$wc_i=[w_j\cdots w_{j+k}]$，$tc_i=[t_j\cdots t_{j+k}]$，$1\leqslant k,j+k\leqslant n$。

我们结合两种特征来产生候选的标注实体，采用的特征模型如下所示：

$$WC^*=\arg\max_{WC} P(WC)\times P(W|WC) \tag{6}$$

$$TC^*=\arg\max_{WC} P(TC)\times P(T|TC) \tag{7}$$

结合式(6)和式(7)可以得出：

$$\begin{aligned}(WC^*,TC^*)&=\arg\max_{(WC,TC)} P(WC,TC|W,T)\\&=\arg\max_{(WC,TC)} P(WC,TC,W,T)/P(W,T)\\&\approx\arg\max_{(WC,TC)} P(WC,W)\times[P(TC,T)]^{\beta}\\&\approx\arg\max_{(WC,TC)} P(WC)\times P(W|WC)\times\\&\quad[P(TC)\times P(T|TC)]^{\beta}\end{aligned} \tag{8}$$

其中，β 是平衡因子，用来调节词形特征和词性特征的权重，避免过度依赖某一特征，$\beta>0$。

在分词的基础上，通过词形特征和词性特征来进行军事语料实体的识别和词性标注的过程，实际上就是对部分词语进行拆分重组和重新确定实体类别，最后确定一个最优的词形或词性序列。

3.1.3 组合特征

从军事用语的角度来看，军事实体并不会存在单一的、固定的称呼，尤其是军事装备的称呼，通常包含多种表达方式，一般说来，军事装备名称可以同时包含字母、短横线、数字和文字这些内容，而且这些内容的组合形式也并不固定，可以用字母＋短横线＋数字表达，也可以用字母＋数字表达，甚至是直接用数字＋文字的方式来表达。因此，建立这种组合形式多样的实体识别特征，对于正确标注军事语料中的实体来说，有着极其重要的作用。

对于这类军事实体，我们采用正则表达式的方法进行处理，军事文本中无论是用语还是实体都有着与其对应的军事特征和军事规则，例如对于“首长”一词的表述，可以用“姓氏”＋“首长”，也可以用“首长”＋“同志”来表述；“HK MP5 冲锋枪”可以用“MP5”和“MP5 冲锋枪”来表示。显而易见，对于这种组合特征，如果仅仅将字符、数字、文字和短横线的组合作为规则标准，那么会造成很多非军事语料实体的误判，从而造成正确率和召回率偏低。在军事语料中，会存在大量表示军事实体的缩略语，尤其是从网络来源获得的军事语料。对于这类词语，传统的方法主要两个：基于词库对比的方式和基于规则统计的方法。Xie 等[11]利用内部结构规则和条件随机场模型，来抽取缩略语和其对应原短语对，完成对缩略语对的识别和抽取；刘群、崔世起[12]通过建立语言模型对候选缩略词集合实现与源短语的对齐，从而得到缩略词典；Change 等[13-14]通过 HMM 模型和内部构成规则进行抽取和识别工作。

为了解决上述组合词语和缩略语的问题，我们采用基于词典的方法和基于规则统计的方法相结合的方式，利用建立好的军语词典，将其中的长词进行切分，然后将切分后的所有字符、数字、和实体短词作为一个组合库内容，建立起这些组合库之间的组合规则，这样就避免了其他字符、数字和文字组合被误判为军事语料实体的问题，并将其纳入特征库中，作为军事语料实体识别的一个重要特征。

3.1.4 军语词典的建立

军事语料中存在很多与军事相关的各种实体，如军职军衔、军事装备、物资设施、军事机构名称(部队番号等)、军用地名等。对这些军事语料实体进行标注，需要以军事领域的知识为依据，而且标注人员的主观认知和知识程度的不同，很容易导致同一军事语料实体被标注为不同的实体或被标注成多个实体。因此，需要建立军事语料实体的词表，将这些词表汇总制作成军事语料实体词典，一方面为实体标注提供参考标准，另一方面也可以将词表引入实体识别中，以提高语料中实体的识别率。

《军语》中包含了军队各项行动和工作中统一使用的军事术语，是一种规范化的标准的军事用语，我们从中提取并建立军事语料实体词表。此外，还需要人工收集军队军职军衔名称、军事装备名称、军用物资名称、军事设施名称、军事机构等、分别构建人员军职军衔词条、军事装备词条、军用物资词条、军事设施词条、军事机构词条等。具体的词条构建数目如表 6 所列。

表 6 军事语料实体词条

词表名称	词条数目	e.g.
军职军衔词条 PER	156	参谋，旅长
军事装备词条 EQU	766	运-20 运输机，歼 16
军用物资词条 MAT	76	被装，单兵干粮
军事设施词条 FAC	69	避难所，军事基地
军事机构词条 ORG	127	总参，军科院
军事地名词条 LOC	37	训练场，西南高地

3.2 自动扩展的军事语料实体特征库的建立

本文设计了一种能够随着训练语料的迭代分类训练自动扩展军事语料实体特征的特征提取框架，首先针对预处理和规范化处理之后的训练语料，利用军事实体词典中提取的典型特征对军事语料中的实体进行高精分类识别，然后再根据分类之后的军事实体集合和非军事实体集合进行基本特征的选择和提取，将两个集合共同的特征去除，并将新的特征加入到特征库中，不确定集合中的军事实体再通过军事实体词典

进行最大正项匹配来实现校正。如此迭代循环上述过程，不断提取新的基本特征加入特征库中，直到没有新的特征出现为止，最后根据最终的特征库建立特征空间来进行语料实体的识别，并对军事语料进行标注。

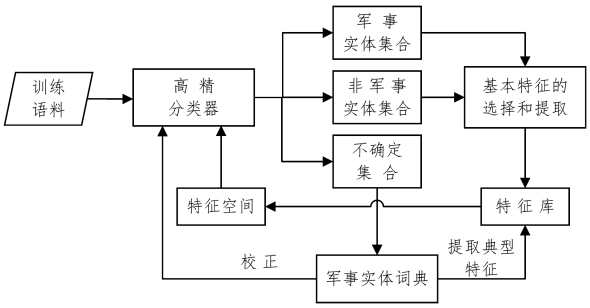

图 2　自动扩展的军事语料实体特征的结构图

最终经过统计和分析可知，军事语料文本中的实体特征和普通的词语特征有着较大的区别，这点在词性特征上显得尤为突出，对于通用领域上的汉语词性标注来说，很多军事词语只能被统称为名词和机构词等，但是对于军事领域来说，这类词语实际上包含了很多细致的类别，不能泛泛地用名词这一词性来标注。此外，军事语料文本中的实体在组合特征上也显得非常明显，在军事领域上的军事语料实体较之通用领域实体具有非常鲜明的组合特征。

3.3　高精分类器的设计

对于军事语料实体的判定实际上是一个分类问题，根据提取的军事语料实体特征进行条件判定，可以将具有相同特征或者特征相近的语料实体，归为军事实体这类实体中。为此，我们建设一个高精分类器，根据 S_k 句子分词之后形成的词组序列 $W_k=\{w_1,w_2,\cdots,w_i,\cdots,w_n\}$，对每个词的特征集 f 赋予一个权值 v，越是典型的军事实体，其特征权值则越大，再将每个词组和其特征权值组合起来构成一个二元组集合 $F_i=\{(f_1,v_1),(f_2,v_2),\cdots,(f_m,v_m)\}$，那么对军事语料中实体的判断算法如算法 1 所示。

算法 1

Input: Sentence S_k, Words W_k, Features F_i, Threshold
Output: Words W_k Foul Score and Foul Detection
1. Foul(W_k)=0;
2. get(f_i)=v_i
3. for i=1 to m
4.　if IsExist(S_k, f_i)=True
5.　then Foul(S_k)=Foul(S_k)+get(f_i)
6.　end if
7. end for
8. if Foul(W_k)>Threshold
9. then
10.　W_k is classified foul.
11. else if Foul(W_k)=0
12. then
13.　W_k is classified not foul.
14. else
15.　W_k is classified uncertainty.
16. end if

通过该算法实现对军事文本的语料实体的分类识别，将各类实体划分为军事实体类和非军事实体类，对于模糊类别的实体暂时将其划为不确定类，在特征库构建完全之后和分类器迭代训练之后，再进行识别分类处理。

3.4　特征选择及校正

在特征的提取过程中，我们在特征向量空间模型中加入了许多特征，如从字、词语、词形和词性等角度提取特征，甚至将英文字符、短横线积极数字的组合作为一种军事实体识别的特征，即能有效解决复杂军事语料实体识别问题，又能丰富并健全军事实体特征集。在如此多的特征中，如何选取合适的特征，或者各种特征应该赋予多大的权值 v，对我们语料库构建中军事语料实体的识别有着重要的影响。在特征的选择中，先用特征统计去除我们建立军事词典提取到的典型军事实体特征，然后对剩余的特征进行频度排序，分别选取其中前 10%，5%，3%，2%和 1%，接着用 x^2 检验对余下的特征进行相关联程度计算，对计算结果进行排序，人工对前 50 个特征进行筛选并确定权值，如此对每份军事语料反复迭代 6 次。x^2 检验特征与分类关系示意图如表 7 所列。

表 7　x^2 检验特征与类关系示意图

特征项 \ 类别	C_j	$\sim C_j$
t_i	A	B
$\sim t_i$	C	D

卡方 x^2 检验是以 x^2 分布为基础的一种常用假设检验方法，它假设 t_i 和类别 C_j 之间符合具有一阶自由度的 x^2 分布，这是统计方法中最为常用的方法之一，公式如下：

$$x^2(F_i,C_j)=\frac{N\times(A\times D-C\times B)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

其中，N 表示军事语料中所有的实体总数，C_j 表示{军事实体}。F_i 表示我们加入的特征，A 表示属于 C_j 类且包含 F_i 的实体频数，即军事语料实体频数，B 表示不属于 C_j 类但是包含 F_i 的实体频数，即非军事语料实体具有军事语料实体特定的特征 F_i，C 表示属于 C_j 类但是不包含 F_i 的军事实体频数，即未被统计的提取出军事实体的语料实体，D 表示既不属于 C_j 也不包含 F_i 的军事实体频数，即非军事语料实体。

4　语料库的实现

我们将收集到的 1040 份军事文本作为军事语料，其中军事新闻 400 篇，军事评论 300 篇，经典战例 150 篇，军事著作 30 篇，军报杂志 120 篇，演习报告 40 篇，共计 1 457 366 字，详细的军事语料信息如表 8 和表 9 所列。将其预处理之后，作为我们的生语料。将其平均等比分成 10 份，其中 9 份语料作为训练语料，用以迭代提取军事语料实体特征，最后一份作为测试语料。

表 8　军事语料的详细信息表

语料来源	字数	词数	句子数	篇章数
军事新闻	267 134	187 803	9 921	400
军事评论	504 144	131 122	7 111	300
经典战例	174 692	322 989	17 711	150
军事著作	227 168	159 705	8 437	30
军报杂志	177 751	128 236	6 611	120
演习报告	106 477	68 471	7 482	40
总计	1457366	998 326	57 273	1040

表 9 军事实体数量统计

军事实体类别	数量
军职军衔名	39745
军事装备名	26947
军用物资名	13958
军事设施名	3205
军事机构名	10968
军用地名	5047
总计	99870

本实验将提取的特征作为分类特征，利用 SVM 对训练语料进行训练后分类，实验结果如表 10 所列。并用 3 个衡量指标来评价最终的语料实体识别结果，即正确率(P)、召回率(R)和 F 值。其计算公式如下：

$$P=\frac{\text{正确识别的军事实体个数}}{\text{识别的全部军事实体个数}}\times 100\%$$

$$R=\frac{\text{正确识别的军事实体个数}}{\text{军事文本中军事实体总数}}\times 100\%$$

$$F=\frac{2\times P\times R}{P+R}\times 100\%$$

表 10 提取不同数量特征的分类结果

序号	特征数	实体总数	识别个数	正确个数	正确率/%	召回率/%	F 值/%
1	300	3216	3119	2736	87.72	85.07	86.38
2	600	3192	3102	2745	88.49	86.00	87.23
3	900	3267	3164	2836	89.63	86.81	88.20
4	1200	3104	3006	2714	90.29	87.44	88.84
5	1500	2983	2897	2640	91.13	88.50	89.80
6	1800	3145	3054	2822	92.40	89.73	91.05
7	2100	3063	2968	2763	93.09	90.21	91.63
8	3000	21970	21381	20498	95.87	93.30	94.57

通过表 10 可以看出，随着我们迭代提取加入特征库中的特征数目的增加，识别的正确率、召回率和 F 值都有所提升，在我们将特征数目提取到 1000 以内时，正确率、召回率和 F 值的最高值未达到 90%以上，没有达到人工识别标注的准确率，换句话说：其标注效果是没有达到能够支撑军事应用的标准。但是当我们继续迭代提取特征，将特征数目扩展到 1500 以上时，准确率、召回率和 F 值都得到了很好的效果，基本满足军事应用的标准。尤其是在我们将 7 份训练语料分别进行完迭代训练之后，再将 7 份语料整合到一起作为一份训练语料进行迭代训练时，特征数扩展到了 3000，最终的准确率、召回率和 F 值的数值达到了 95.87%，93.30%和 94.57%，达到了较高的水平，满足军事应用的标准。

此外，未在表 10 中列出的还有一点值得指出，在我们采集的军事语料中，对于专业的军事文本(如军事文书)的语料实体识别标注，其准确率始终高于来源于网络的军事文本语料，主要原因就是在语料的规范化处理上，无论对网络来源的军事语料进行什么样的规范处理，也无法得到专业的军事语料的规范和标准效果，无论是从用语上还是文法上，专业军事文本语料都十分注重这方面，而网络军事文本语料大多存在一些口语化和网络化的描述方式，因此造成了网络军事语料实体被识别和标注的效果较低。

结束语 本文对军事语料中的实体标注进行了介绍和分析，提出了军事语料标注规范并设计了一种军事词语词性标记规范，设计了一种多次迭代自动扩展的军事语料实体特征库的方法，对军语实体特征进行选择和提取，构建军事语料实体特征空间，识别军事语料中的实体并进行语料标注，构建了一个较高质量的面向实体的军事语料库。以军事语料库为基础，结合相关系统，不仅可以进行军事语言的研究学习，还能提高信息获取能力，给我军节一个在军事行动中可依托的语料库，大大提升了我军信息化作战的战斗力和生存能力，具有较大的应用前景和研究意义。

参 考 文 献

[1] 麻丽莉，王祥兵. 军事平行语料库的建立及其在军事翻译方面的应用[J]. 国防科技，2009，30(1)：38-41.

[2] 梁晓波，刘伍颖，孟凡礼. 信息化条件下的军事语料库应用[J]. 国防科技，2008(2)：51-57.

[3] 王红霞，周密. 国际化视域下海军军事科技英语的实用性研究[J]. 中国校外教育旬刊，2014(S1)：1103-1104.

[4] 向音. 军用文书的语篇特征初探[J]. 办公室业务，2011(10)：010.

[5] 俞士汶，朱学锋，段慧明. 大规模现代汉语标注语料库的加工规范[J]. 中文信息学报，2000，14(6)：58-64.

[6] 范云飞. 基于 POS 规则匹配的电子商务网站用户评价信息的分析[D]. 武汉：武汉理工大学，2015.

[7] XIA F，YETISGEN-YILDIZ M. Clinical corpus annotation：Challenges and strategies[C]// Proceedings of the 3rd Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) of the International Conference on Language Resources and Evaluation (LREC). 2012：32-39.

[8] SNOW R，O'CONNOR B，JURAFSKY D，et al. Cheap and fast—But is it good? Evaluating non-expert annotations for natural language tasks[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg. Association for Computational Linguistics，2008：254-263.

[9] ZHOU J，LI B C，CHEN G. Automatically building large-scale named entity recognition corpora from Chinese Wikipedia[J]. Frontiers of Information Technology & Electronic Engineering，2015，16(11)：940-957.

[10] NADEAU D，SEKINE S. A survey of named entity recognition and classification[J]. Lingvisticae Investigations，2007，30(1)：3-26.

[11] XIE L，ZHENG Y，LIU Z，et al. Extracting Chinese abbreviation-definition pairs from anchor texts[C]// International Conference on Machine Learning and Cybernetics. IEEE，2011：1485-1491.

[12] 崔世起. 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). 2005：6.

[13] CHANG J S，TENG W L. Mining atomic Chinese abbreviations with a probabilistic single character recovery model[J]. Language Resources and Evaluation，2007，40(3-4)：367-374.

[14] CHANG J S，LAI Y T. A Preliminary Study on Probabilistic Models for Chinese Abbreviations[C]// Proceedings of the Third Sighan Workshop on Chinese Language Learning. 2004：9-16.