文章编号：1003-0077(2020)08-0041-10

面向国防科技领域的技术和术语语料库构建方法

冯鸾鸾，李军辉，李培峰，朱巧明

（苏州大学 计算机科学与技术学院，江苏 苏州 215006）

摘 要：互联网存在海量的文献和科技信息，隐含着大量高价值情报。识别国防科技领域中的技术和术语可以为构建国防科技知识图谱奠定基础。该文基于此领域的海量军事文本，以维基百科中军事领域的新技术为基点采集语料，涵盖了新闻、文献和维基百科三种体裁。在分析军事技术文本特点的基础上制定了一系列标注规范，开展了大规模语料的标注工作，构建了一个面向国防科技领域的技术和术语语料库。该语料库共标注了 479 篇文章，包含 24 487 个句子和 33 756 个技术和术语。同时，该文探讨了模型预标注策略的可行性，并对技术和术语类别在不同体裁上的分布以及语料标注的一致性进行了统计分析。基于该语料库的实验表明，技术和术语识别性能 F_1 值达到 70.40%，为进一步的技术和术语识别研究提供了基础。

关键词：面向国防科技领域；技术和术语；标注规范；语料库

中图分类号：TP391 **文献标识码**：A

Constructing a Technology and Terminology Corpus Oriented National Defense Science

FENG Luanluan, LI Junhui, LI Peifeng, ZHU Qiaoming

(School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China)

Abstract: Massive literature and science information on Internet can supply valuable intelligence. The detection of technology and terminology is fundamental for constructing oriented national defense science (ONDS) technology knowledge base. We analyze military text characteristics and design annotation guidelines for ONDS technology and terminology from massive internet content for a list of military emerging technology defined in Wikipedia. Based on the annotation guidelines, we conduct broad-scale corpus annotation process, and we construct a ONDS technology and terminology corpus which covers three genres of news, papers and Wikipedia. we finally annotated 479 articles with 24,487 sentences and 33,756 technologies and terminologies. Meanwhile, we explore the feasibility of model pre-annotating, analyze distribution of technology and terminology in different genres and calculate annotation consistency for the corpus. Experiment result based on the corpus show that the detection of technology and terminology achieves 70.40% F_1 scores. The work presented in this paper builds foundations for detection of ONDS technology and terminology.

Keywords: oriented national defense science; technology and terminology; annotation guidelines; corpus

0 引言

随着互联网信息的不断增长，从大数据中挖掘有价值的信息并应用于国防建设是一个必然趋势。在互联网上存在海量的文献和科技信息，从中可以得到高价值情报。研究如何从大数据中抽取世界各国国防相关技术及其研发和应用信息，有助于把握国防技术发展的态势，为我国国防建设服务。

传统意义上的实体通常指人名、地名和机构名称。显然，在国防科技领域，技术和术语不同于传统意义上的实体。例如，在例 1 中，“Synthetic aperture radar”及其缩略语“SAR”都可以认为是特殊类别的实体，即属于本文定义的技术类别。“airborne SAR system”是一个名词短语，同时也被视为一项技术。又如例 2 中“1995 United Nations Protocol

收稿日期：2019-09-16 **定稿日期**：2019-10-09

基金项目：国家自然科学基金(61836007,61472265,61876120)

on Blinding Laser Weapons"这一协议,在本文中视作一个特定军事术语。由这些例子可以发现,面向国防科技领域的技术和术语通常很长,有的可能同时包含形容词和连词,并且一个技术或术语可能会有多种不同的表现形式,比如简写或者首字母缩略词等。

例1 *The Synthetic aperture radar (SAR) technology has been developed in China since 1970s, and the first airborne SAR system was established in 1979 and obtained multiple SAR images.*

例2 *Blinding laser weapons have been tested in the past, but were banned under the* ***1995 United Nations Protocol on Blinding Laser Weapons.***

目前命名实体识别技术日臻成熟,在信息抽取、机器翻译、自动问答等领域有着广泛应用,然而已有的国内外研究大多面向通用领域。因此针对国防科技领域技术和术语的独特性,参考美国国防军事术语词典①,本文首先建立了包括基础技术、综合技术、武器、组织以及军事术语五大类的面向国防科技领域的技术和术语标注体系,示例如表1所示,并制定了相应的标注规范,最终构建了具有一定规模的面向国防科技领域的技术和术语语料库,为后续技术和术语识别研究打下基础。

表1 面向国防科技领域技术和术语分类示例

类别	示例
基础技术(基础技术是存在于自然界中、人们可以加以改造应用的通用领域的科学理论和研究方法)	kinetic energy(动能) non-ionising radiation(非电离辐射)
综合技术(国防科技领域内能够直接为人类所用的工具、方法等)	active electronically scanned array(有源电子扫描阵列) low-probability-of-intercept radar(低截获概率雷达)
武器(综合技术的具体实例或应用,标注具体型号或名称)	Glare LA-9/P(激光眩目器的一种型号) Avro Vulcan(火神轰炸机)
组织(与综合技术相关的公司、军队等)	United States Military(美国军方) DARPA(美国国防高级研究计划局)
军事术语(综合技术涉及到的项目、行动等,在介绍综合技术性能时用到的专有名词等)	World War II(第二次世界大战) 1995 United Nations Protocol on Blinding Laser Weapons (1995年关于激光致盲武器的议定书)

1 相关工作

近年来,语料库构建的研究引起了国内外学者的广泛关注。目前构建语料库的方法有自动构建[1-2]和人工构建[3-5]两种,在对话[6]、语言[7-9]、微博[10-11]、医学[12]等领域均已出现公开构建的语料库,并得到广泛应用。Lowe等[6]基于Ubuntu社区的对话内容,构建了Ubuntu对话语料库,其中包含一百万个对话内容。该语料既有Dialog State Tracking Challenge数据集的多次序对话特性,也有类似Twitter上的人类自然对话特点,已成为对话系统的公开评测数据集。Jiang[13]等选取CTB8.0中的所有新闻语料,结合宏观篇章理论和修辞结构理论,标注了包含720篇文章的中文宏观篇章语料库。Chen和Nie[14]基于爬取技术,爬取双语平行的网页内容,构建了跨语言中英平行语料库,包含117.2MB的中文文本和136.5MB的英文文本。Hu[10]等构建了取自于新浪微博的大规模中文短文本摘要数据集,其中包含了超过200万个真实的中文短文本数据和每个文本作者给出的简短摘要。Peng[15]等基于中文微博信息,构建了第一个中文社交媒体语料库,其中包含人名、机构名、地名和地缘政治四个命名实体类型,而且每个类型都包括特定实体和指代实体。杨锦锋等[16]收集医学领域的中文电子病历数据,结合中文病历特点,制定了命名实体和实体关系标注体系,构建了包含992份病历文本的中文电子病历命名实体和实体关系语料库。由于这些语料库具有领域特性,所以很难移植到国防科技领域。而与国防科技领域相近的军事领域大多针对的是部队、机构、军用地名等军事命名实体识别研究[17-18],目前尚未发现公开的面向国防科技领域技术和术语语料库。因此,需要构建此类语料库以进行面向国防科技领域技术和术语识别的深入研究。

① http://www.jcs.mil/Doctrine/DOD-Terminology/

2　面向国防科技领域技术和术语标注规范的制定

2.1　标注体系

近年来，与国防科技领域相近的军事领域的命名实体识别研究越来越受到重视，因此有一些小规模的军事实体数据集。例如，单赫源等[17]对作战文书中的部队编制、作战编成、装备型号和任务等进行了标注；冯蕴天等[18]在收集到的军事文本中标注出军事装备名、物资名、设施名和机构名等各种与军事相关的命名实体；王学锋等[19]收集并标注了军事相关语料集，其中包括机构、地名、武器、时间等军事命名实体。本文参考上述文献，并结合技术和术语识别任务的特点制定标注体系，技术和术语识别是构建技术知识图谱的基础工作，而一个技术知识图谱，需要包含技术的属性和技术的层次关系，技术中又包括通用类的基础技术。因此为了后续工作的较好开展，本文将技术和术语更加细致地划分为五类，以综合技术为核心，围绕综合技术标注与其相关的基础技术、武器、组织和军事术语。其中综合技术是指面向国防科技领域的技术，是本文所采集的军事技术文本中的核心；基础技术指通用领域技术；武器则扩展为综合技术的实例和综合技术的应用；组织指与综合技术相关的公司、军队等；最后，军事术语指综合技术涉及到的项目、行动等。下面将详细描述这五类技术和术语的相关特征与示例。本文所定义的技术和术语遵循不重叠、不嵌套标注原则。

2.1.1　基础技术

基础技术是存在于自然界中、人们可以加以改造应用的通用领域的科学理论和研究方法，本文指的是采集的军事技术文本中在讲述综合技术原理时涉及到的一些物理科学和理论等。包括以下几个方面。

(1) 物理技术，如电磁频谱、无线电波等。

Multiple laser beams with non-overlapping optical spectra(非重叠光谱) are combined……

(2) 物理理论，如动能、电能、多普勒效应等。

The energy from the expansion of gases on firing appears in the form of kinetic energy(动能) transmitted to the bolt mechanism.

(3) 物理或化学材料。

These are copper nickel-coated glass fibers(铜镀镍玻璃纤维) or silver-coated nylon fibers(镀银尼龙纤维) having lengths equal to half of the anticipated radar wavelength.

2.1.2　综合技术

综合技术为本文建立的标注体系的核心，指国防科技领域内能够直接为人类所用的工具、方法等。军事技术文本围绕一个核心的综合技术展开，涉及多个综合技术，具体可分为以下四类。

(1) 能够直接为人类所用的工具、方法等。

The principal radar technology being introduced is active electronically scanned array (AESA)(有源电子扫描阵列) technology.

(2) 直接使用基础技术的弹药等。

Even advanced kinetic energy ammunition(动能弹药) such as the United States' M-829A3 is considered only an interim solution against future threats.

(3) 技术类武器的集合。

Directed-energy weapons(定向能武器) are still very much at the experimental stage and……

(4) 技术集成的系统。

Another important application of high-energy directed energy laser systems(高功率定向能激光系统) when operated at relatively-lower power levels ……

2.1.3　武器

在标注的军事技术文本中，武器是某种综合技术的具体实例，例如，“AAM-4B”是“air-to-air missile”(空对空导弹)的具体实例、M551 Sheridan 是轻型坦克的一种。另一方面，武器也可以是综合技术的应用，如“AESA”(有源电子扫描阵列)应用在“AAM-4B”上、“EMALS”(电磁飞机发射系统)应用在“Queen Elizabeth-class”上。为了方便标注，将武器简单分为以下三类。在标注时必须标注具体型号或名称。

(1) 特定型号的枪支弹药、战舰、导弹等。

The first AESA(有源电子扫描阵列) on a missile is the seeker head for the **AAM-4B** air-to-air missile(空对空导弹).

(2) 特定名称的枪支弹药、战舰、导弹等。

Intended for use on the **M551 Sheridan** light tank(轻型坦克), the **Shillelagh** missile was fired out of the **Sheridan**'s cannon to provide robust anti-tank capability.

(3) 特定名称+class表示战舰。

A contract was signed in December 2011 with General Atomics of San Diego to develop EMALS (电磁飞机发射系统) for the **Queen Elizabeth-class** carriers.

2.1.4 组织

在本文建立的标注体系中,将组织局限于与综合技术相关的两个方面,如下所示:

(1) 研发综合技术或生产综合技术产品的公司、实验室。

LE Systems makes the Laser Dazzler(激光眩目器), which resembles an ordinary torch and emits a low power pulsing green laser light.

(2) 综合技术或综合技术产品为其提供服务的海军、空军、战队等。

The GATS (GPS Aided Targeting System) (GPS辅助定位系统) on the **USAF(美国空军)** B-2A is a good example of such a system.

2.1.5 军事术语

军事技术文本中的军事术语大多与综合技术相关,在介绍综合技术的发展历程时会涉及很多项目、行动等,另外还有一些在介绍综合技术性能时用到的专有名词等。本文将军事术语分为以下四个方面。

(1) 与综合技术相关的协议、特性。

Blinding laser weapons(致盲激光武器) have been tested in the past, but were banned under the **1995 United Nations Protocol on Blinding Laser Weapons(1995年关于激光致盲武器的议定书)**.

(2) 综合技术应用的战争。

The 30 mm MK 108 cannon was perhaps the apogee of API blowback technology(API反冲技术) during **World War Ⅱ(第二次世界大战)**.

(3) 名称指代的项目、工程、系统、行动等专有词汇。

This device was known as the **Phoenix project (凤凰工程)** within the **Strategic Defense Initiative research program(战略防御倡议研究计划)**.

(4) 特殊标注示例:非首字母大写的专有名词。

Accuracy is expressed as **circular error probable (CEP)(圆概率误差)**.

2.2 标注准则

语言的表达方式多样化,而且语料中常常含有大量具有争议或模棱两可的语言现象,使得标注者往往难以达成一致。因此,需要定义一套标注准则以规范争议性实例的标注,提高语料的标注质量,同时降低标注难度。根据语料和任务特点,本文需要规范技术和术语的边界标注问题,降低标注的不一致性,使得构建的语料库有助于后续技术知识图谱构建的实际应用。具体的标注准则如下:

(1) 标注最大范围

例3 *MICA-IR, manufactured by* MBDA, *is a short - and medium-range air-to-air missile having a maximum operational range of 50 km and a maximum speed of 3 Mach.*

例3中"short- and medium-range"为"air-to-air missile"(空到空导弹)的修饰词,是"air-to-air missile"的一个类别,遵循最大范围标注,遇到综合技术前面有修饰词表明综合技术的类别,将其作为综合技术的一部分一起标注。

(2) 遵循特异性标注

例4 *For the DDG 51 Flight Ⅲ destroyer, the* ***SPY-6*** **(V)** *AMDR will feature……*

例4中"SPY-6(V)"是"AMDR"(防空反导雷达)的应用实例,将其标注为武器类,"AMDR"标注为综合技术类。即在最大范围内根据特异性分别标注出技术、术语类别。

(3) 不标注特殊符号

例5 *Electrothermal-chemical* (ETC) *technology is an attempt to increase accuracy and muzzle energy of future tank.*

例5中"technology"被括号隔开,此时不标注"technology",括号中"ETC"为前面"Electrothermal-chemical"的缩写,两者分别标注为综合技术。

(4) 单个常见技术不标注

例6 APAR*'s missile guidance capability supports the Evolved Sea Sparrow Missile* (ESSM) *and the* **SM-2 Block IIIA** *missile.*

例6中"SM-2 Block IIIA missile"将"missile"前面的型号标注为武器,单个"missile"不标注。同样,单个"tank""aircraft"等都不标注。

(5) 维基百科作为参考,避免臆断推测

在标注时,因为军事专业知识的局限,很容易漏标和标错。因此,对于不确定的技术和术语,以维基百科作为参考,查询其意,避免主观臆断。

3　面向国防科技领域技术和术语语料库的构建

3.1　语料的选取与预处理

本文构建的国防科技领域技术和术语语料库覆盖新闻、文献和维基百科三个体裁，其中新闻类语料选自各大军事新闻网站①，文献类语料主要选自IEEE② 网站。在文档采集时，以维基百科中的军事领域新技术③为基点，以综合技术为中心采集语料。这样能够避免标注信息过于分散和稀疏，且保证各项新技术语料数量的均衡性。

人工采集文档时会剔除文章中不含有技术和术语的内容，保证语料的质量。人工采集的语料保存为文本格式。标注之前对数据进行符号化处理，即将标点符号与单词隔开，同时删除空行和去掉多余的空格。

标注格式如图 1 所示，标注文本中[@... #.. *]表示一个技术和术语，#.. * 间表示技术和术语的类别。标注工具的匹配功能会将相同的技术和术语标注出来作为推荐，减轻标注人员负担，标注时只需检查推荐部分。

> The system is connected to ***[@shipboard missiles#DTEC*]*** and ***[@machine guns#DTEC*]*** for engaging incoming threats .
>
> WASHINGTON – ***[@U.S. Navy#ORG*]*** missile-defense experts are asking the ***[@DRS Laurel Technologies#ORG*]*** segment to provide ***[@AN/SPQ -9 B#WEA*][@anti-ship missile defense#DTEC*]*** (***[@ASMD#DTEC*]***) radar systems to help protect ***[@U.S. Navy#ORG*]*** surface warships from enemy ***[@anti-ship missiles#DTEC*]*** .

图 1　技术和术语标注语料示例

3.2　标注过程

本文参照中文电子病历命名实体和实体关系语料库[16]的标注过程进行标注。首先，考虑到标注的文本内容具有较强的军事专业性，在试标注阶段采用反复标注并讨论的策略制定标注规范，在正式标注阶段，制定了多轮标注策略。

3.2.1　试标注阶段

首先由两名规范制定者分析军事科技文本特点，对比与其他军事命名实体定义的差别，并与两位军事技术专家多次交流，确立了技术和术语的分类框架，制定了初步的标注规范。以此初步规范分别独立试标注 10 篇军事技术文档，计算一致性并分析不一致的标注，不断讨论并参考军事技术专家意见更新标注规范。之后再以新的标注规范重复试标注步骤以完善标注规范，直到两人的标注一致性达到 80％以上。从以下三个方面保证规范的质量和可操作性：

（1）对每一类技术和术语的标注，给出相应的定义，并更细致地总结出从哪些方面入手，并辅以示例，更有益于理解和标注；

（2）针对一些容易混淆的问题，规范列出大量特殊标注示例，并制定标注准则；

（3）根据标注规范标注的文档一致性达到 80％以上，保证了规范的质量。

3.2.2　正式标注阶段

制定了标注规范后，经过培训筛选，最终确认 18 名标注人员（包括两名标注规范制定者）。为了获取大规模标注文本，制定了多轮标注策略。流程如下：

（1）前两轮标注实行三标注，标注人员分为 6 组，每组 3 名标注人员分别独立标注同一篇文章，最后统计 3 名标注人员的标注一致性，3 名标注人员讨论标注语料中不一致的标注，征求专家意见以统一标注语料中不一致的标注；

（2）第三轮挑选出标注质量高的 6 名标注人员，分为三组进行二标注，即由两名标注人员分别独立标注同一篇文章，由规范制定者进行审核，并与两名标注人员进行讨论，统一两份标注语料中不一致的标注。

为了保证组间一致性，每轮会重新对标注人员进行分组。最终形成面向国防科技领域的技术和术语语料库。具体标注过程如图 2 所示。

3.3　模型预标注探讨

考虑到人工标注需要消耗大量人力物力，本文探索了模型预标注方案，期望该方案能够在不降低标注质量的条件下，节省人工标注时间。为此，在标注了 200 篇文档后，用这 200 篇文档训练了一个模

① https://www.militaryaerospace.com，https://spacenews.com 等

② https://ieeexplore.ieee.org/Xplore/home.jsp

③ https://en.wikipedia.org/wiki/List_of_emerging_technologies#Military

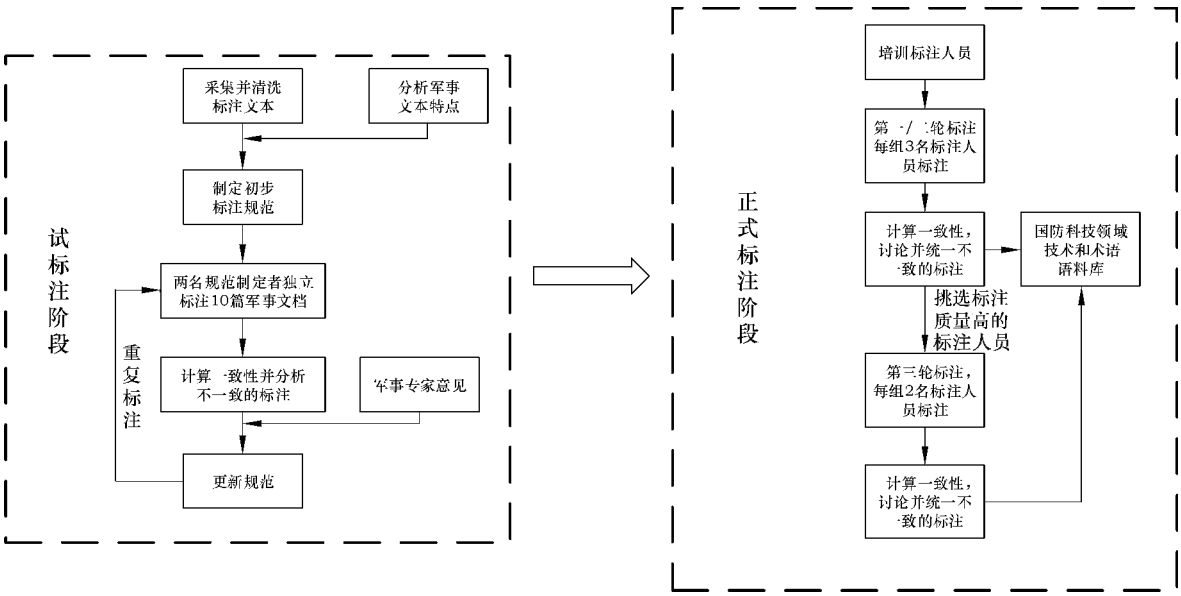

图2 语料标注过程

型以进行模型预标注，即人工在模型预标注的结果上进行修正。选择10篇文档让两名标注人员分别进行独立标注，其中一名(A)直接标注生文档，另一名(B)在模型预标注的结果上进行修正。由规范制定者进行审核，并与两名标注人员进行讨论，统一两份标注语料中不一致的标注，形成最终语料。分别计算两份独立标注的语料间的一致性及其与最终语料间的一致性。为规避偶然性，三组标注人员均进行上述标注试验。一致性结果如表2所示。

表2 标注试验一致性

组别	标注人员	独立标注一致性/%	与最终标注一致性/%	标注时间/篇
第1组	A	71.1	88.1	30min
	B		80.7	20min
第2组	A	73.8	89.8	25min
	B		81.3	18min
第3组	A	74.1	88.7	33min
	B		83.7	25min

由表2可以看出，与原期望不一致。在模型预标注结果的基础上进行修正标注的确可以节省一部分时间，但相应的标注质量却下降了。分析有以下原因：

(1) 标注人员在有了参考答案的基础上通过前期标注训练获得的标注能力会逐渐退化，会慢慢忘记自己的那套标注系统，转而过分依赖于模型训练得到的结果；

(2) 只用前期标注的200篇语料进行训练得到的模型不够稳定，精确率只有不到68%，在标注过程中会出现很多错误，譬如模型标注“lightweight pulsed-power supply(轻型脉冲电源)”为综合技术，而在规范定义中，lightweight(轻型)这种不含有技术性的修饰不标。因此模型标注的结果会误导标注人员。

随着标注语料的增多，训练的模型质量逐步稳定，模型预标注在节省时间方面是可取的方法。但模型预标注需要达到什么样的性能，才能在不降低标注质量的条件下节省标注时间，仍是后期要探讨的问题。

3.4 语料库统计

3.4.1 标注数量统计

语料库的统计信息如表3所示，覆盖新闻、文献和维基百科三个体裁，共计标注了479篇文章，总计24 487句，标注了33 756个技术术语。

表3 语料库的统计信息

统计条目	新闻	文献	维基百科	总数
篇章数	179	199	101	479
句子数	8 444	8 665	7 378	24 487
技术和术语个数	12 461	10 885	10 410	33 756

图3展示了语料中各类别技术和术语数量统计，因为本文围绕综合技术采集文档，因此在标注样例中，综合技术类别技术和术语最多，约占总数的

73.5%。武器作为某些综合技术的具体实例和应用，数量上占总数的 11%。组织和军事术语分别占 6%和 7%。基础技术因为较少提及，数量最少，只占总数的 2.5%左右。

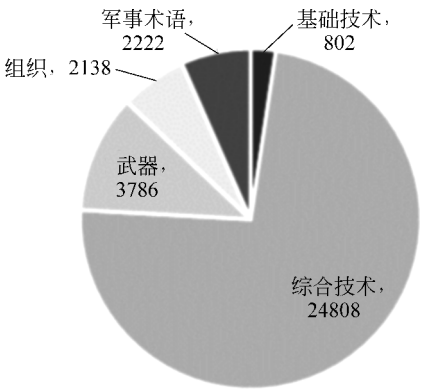

图 3　语料中各类别技术和术语数量统计

语料库覆盖了不同体裁，图 4 展示了技术和术语类别在不同体裁上的分布情况，可以看出总体上，技术和术语类别在维基百科和新闻两体裁上的的分布情况相似，而在文献上的分布与其他两种体裁有较大差异。维基百科和新闻中武器和组织类别占比多于文献，因为维基百科和新闻体裁在提到综合技术时，常常会进一步提到与其相关的组织机构，以及其应用和具体实例，而文献更多地则会阐述关于综合技术的原理，因此基础技术占比多于其他两类。令人感到诧异的是，军事术语类别在文献中的占比多于其他两种体裁。这主要是因为本文定义的军事术语不仅仅限于综合技术涉及的项目、行动等，还包括在介绍综合技术性能时用到的专有名词，因此文献中的军事术语类别占比多于新闻和维基百科。

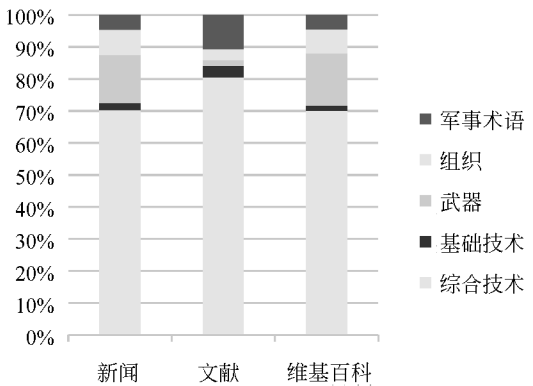

图 4　语料中技术和术语的类别分布

3.4.2　标注一致性分析

标注一致性一般可以用两种指标衡量：Kappa 值[20]和 F_1 值[21]。Kappa 值一般适用于标注类别正例和负例较均衡的场景，比如情感极性分类的语料标注。而在实体识别语料标注中，未标注的文字通常被视为负例，而且在数量上远远多于标注为实体的文字。在该场景下，可以采用 F_1 值评估标注一致性[21]。因此，本文使用 F_1 值评价语料标注的一致性。具体做法是，视一个标注者(A1)的标注结果为标准答案，计算另一标注者(A2)标注结果的准确率(P)和召回率(R)，以及 F_1 值，如式(1)～式(3)所示。在将独立标注语料与最终语料比较时，将最终语料视作 A1，即为标准答案，独立标注语料视为 A2。

$$P = \frac{\text{A1 和 A2 一致的标注结果总数}}{\text{A2 的标注总数}} \tag{1}$$

$$R = \frac{\text{A1 和 A2 一致的标注结果总数}}{\text{A1 的标注总数}} \tag{2}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{3}$$

表 4 分别给出三轮标注一致性的各组平均 F_1 值。第二列“分类”表示两个实体标注一致当且仅当实体边界完全匹配且类别一致；第三列“识别”表示两个实体标注一致当且仅当实体边界完全匹配。各列中的第一项表示独立标注语料间的一致率，第二项表示独立标注语料与最终语料间的一致率。

表 4　技术和术语标注一致性各组平均 F_1 值

次数	分类/%	识别/%
第一轮	71.81/75.69	76.21/79.59
第二轮	75.49/78.39	80.05/82.62
第三轮	75.77/86.49	80.03/88.84

从表 4 中可以看出：

(1) “分类”一致性 F_1 值要低于“识别”一致性 F_1 值。因为技术和术语的标注需要一定的军事背景知识，在识别出相应的技术和术语并将其准确分类上有一定的难度。

(2) 独立标注的语料间的一致性低于它们与最终语料间的一致性。因为最终语料是军事科技专家或规范制定者审核，并与标注人员共同讨论不确定和不同的标注，再进行修改获得的，有和独立标注语料相同的部分。

(3) 第三轮独立标注与最终语料间的一致性要高于第一轮和第二轮。这是因为前两轮都是三标注，即最终语料是由军事科技专家与 3 名标注人员讨论修改的，即综合了三份标注语料的意见，所以独立标注语料与最终语料的一致性不高；第三轮标注

经过前两轮的训练，挑选了标注质量好的标注人员，最终语料由规则制定者整合了两份标注语料的结果，因此一致性会高于前两轮。

3.5 与其他命名实体识别语料库比较

命名实体识别是自然语言处理领域的一个经典问题。研究者常用的语料库包括：

(1) CoNLL03是目前通用领域常用的公开数据集。其英文数据集选自路透社1996年8月至1997年8月的新闻，德文数据集选自ECI多语言文本语料库。CoNLL03定义的命名实体包括人名、地名、机构名和其他四种类型。

(2) Weibo NER是2015年发布的第一个中文社交媒体语料库，语料选自2013年11月至2014年12月期间发布的微博信息，包括人名、机构名、地名和地缘政治四个实体类型，每个类型都包括特定实体和指代实体。

(3) NCBI是在疾病医疗领域广泛应用的英文疾病医疗语料库，语料选自PubMed摘要，包括特殊疾病、疾病类名、复合名和调节剂四个实体类型。

表5给出了本文语料库与目前命名实体识别相关研究所使用的一些语料库的规模比较。可以看出，本语料库的规模与目前通用领域流行的CoNLL03英文语料库相当，在句子数和实体数上均高于其他三个语料，这也从侧面表明了本文构建的语料库适用于面向国防科技领域的技术和术语识别研究。

表5 与通用语料库规模对比

语料库	句子数	实体数	实体类别个数
CoNLL03[22](英文)	22 137	35 099	4
CoNLL03[22](德文)	18 933	17 357	4
Weibo NER[15]	1 890	2 688	4
NCBI[23]	7 856	6 892	4
技术和术语语料库	24 487	33 756	5

3.6 语料库应用分析

从后续应用来看，本文构建的面向国防科技领域技术和术语语料库，在技术关系抽取、技术属性和描述抽取等方面都有应用价值。譬如在关系抽取中，通过此语料库中标注出的技术可以进一步识别出技术间的层次、同义等关系。又如在技术属性和描述抽取方面，很多术语提示了技术的属性所在，通过此语料库可以更加容易地进行这方面的研究。在构建面向国防科技领域的知识图谱方面，自动识别技术和术语是一项最基本的研究内容。

4 技术和术语识别实验

为了说明本文构建的面向国防科技领域技术和术语语料库的可计算性，对技术和术语识别进行了初步的实验。本文使用传统的序列标注模型Bi-LSTM+CRF[24]进行实验，将479篇文章按照5∶1∶1的比例，随机选择344篇作为训练集，68篇作为开发集，67篇作为测试集。首先使用NLTK对数据进行分句处理，然后将数据转换为序列格式。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，我们将报告区分和不区分技术和术语类别的性能。

实验参数设置如表6所示，采用的字符向量维度是30，字符LSTM的隐层大小是50。采用glove预训练的词向量①对词汇进行初始化，词向量维度是100，词LSTM的隐层大小是300。使用随机梯度下降(SGD)算法训练模型，设置一个批次的样本数为10，迭代次数为100，学习率设为0.005，并采用Hinton等人提出的dropout方法将隐层的节点以0.5的概率随机忽略[25]。

表6 实验参数

参数	训练值	参数	训练值
字符向量维度	30	批样本数	10
字符LSTM隐层大小	50	迭代次数	100
词向量维度	100	学习率	0.005
词LSTM隐层大小	300	dropout rate	0.5

各模型的性能表现如表7所示，括号外指区分类别的性能，括号内指不区分类别的性能，其中WLSTM-CRF是基于词的Bi-LSTM+CRF模型，即基准模型，CLSTM-WLSTM-CRF是基于字符的Bi-LSTM+CRF模型。

表7 技术和术语识别实验结果

模型	P/%	R/%	F_1/%
WLSTM-CRF (基准模型)	71.76 (75.89)	66.01 (69.81)	68.76 (72.73)
CLSTM-WLSTM-CRF	73.30 (78.65)	67.72 (72.67)	70.40 (75.54)

① http://nlp.stanford.edu/projects/glove/

实验结果表明，基于字符的模型 F_1 值比基准模型 F_1 值提高 1.64%（2.81%），说明字符序列通过 Bi-LSTM 获得了一些仅用词向量训练抽取不到的上下文信息。同时，评测时是否区分类别对性能的影响为区分类别比不区分类别性能差 4%～5%，表明在识别出相应的技术和术语的基础上再将其准确分类有一定难度，这与我们在标注语料时情况一致。

表 8 给出了基于字符模型的各类别的识别性能。可以看到，基础技术和军事术语的识别性能较低，主要表现在召回率上，综合技术、武器和组织类别的识别效果较好。这与我们在标注语料时情况相似，基础技术数量较少，标注时容易漏标，军事术语容易被标注成其他类别。

表 8　各类别识别结果

类别	P/%	R/%	F_1/%
基础技术	56.92	37.00	44.85
综合技术	73.22	71.80	72.50
武器	71.32	68.27	69.76
组织	82.63	65.22	72.90
军事术语	74.55	35.14	47.77

5　结论与展望

本文描述了面向国防科技领域的技术和术语语料库的构建工作。首先，本文根据国防科技领域的专业知识，通过不断反复试标注，制定了标注军事文本中技术和术语的标注规范。然后基于此标注规范，采用多轮标注策略，展开大规模语料标注，构建了目前规模较大的面向国防科技领域技术和术语语料库，总共标注了 479 篇，共计 24 487 句及 33 756 个技术和术语。同时，本文探讨了模型预标注方法的可行性，并对技术和术语类别在不同体裁上的分布以及语料标注的一致性进行了统计分析。统计表明，技术和术语类别在新闻和维基百科上的分布相似，而在文献上的分布与以上两种体裁有较大差异，语料标注一致性较好。最后，应用该语料库进行初步技术和术语识别实验，实验结果表明，技术和术语识别性能 F_1 值达到 70.40%。

由于面向国防科技领域技术和术语语料库的建设需要人工标注，并且标注时需要不断查阅军事知识，标注有一定的难度，工作量很大，这些制约了本文语料库的规模。未来的工作中，将应用本文语料库进行国防科技领域技术和术语识别的深入研究，并继续完善标注体系，改进标注质量，扩大语料规模，为后续技术归并和层次关系构建、技术属性和描述抽取以及技术知识图谱构建夯实基础。

参考文献

[1] Brockett C, Dolan W B, Dolan B. Support vector machines for paraphrase identification and corpus construction[C]//Proceedings of the 3rd International Workshop on Paraphrasing (IWP2005), 2005.

[2] Dolan B, Brockett C. Automatically constructing a corpus of sentential paraphrases[C]//Proceedings of the 3rd International Workshop on Paraphrasing (IWP2005), 2005.

[3] Vincze V, Szarvas G, Farkas R, et al. The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes [J]. BMC Bioinformatics, 2008, 9(Suppl 11): S9-S9.

[4] Zou B W, Zhu Q M, Zhou G D. Negation and speculation identification in Chinese language[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015: 656-665.

[5] 周惠巍，杨欢，徐俊利，等. 中文模糊限制信息范围语料库的研究与构建[J]. 中文信息学报，2017，31(3)：77-85.

[6] Lowe R, Pow N, Serban I V, et al. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems[EB/OL]. (2016-2-4). https://arxiv.org/pdf/1506.08909v3.pdf.

[7] 奚雪峰，褚晓敏，孙庆英，等. 汉语篇章微观话题结构建模与语料库构建[J]. 计算机研究与发展，2017，54(8)：1833-1852.

[8] Xue N W, Chiou F D, Palmer M. Building a large-scale annotated Chinese corpus[C]//Proceedings of the 19th International Conference on Computational Linguistics, 2002.

[9] Aksan Y, Aksan M, Koltuksuz A, et al. Construction of the Turkish national corpus (TNC)[C]//Proceedings of the 8th International Conference on Language Resources and Evaluation, 2012.

[10] Hu B T, Chen Q C, Zhu F Z. LCSTS: A large scale Chinese short text summarization dataset[EB/OL]. (2016-2-19). https://arxiv.org/pdf/1506.05865.pdf.

[11] Quan C Q, Ren F J. Construction of a blog emotion corpus for Chinese emotional expression analysis

[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009: 1446-1454.

[12] 游正洋，王亚强，舒红平. 基于词性标注的中医症候名语料库[J]. 电子技术与软件工程，2017，21：177-178.

[13] Jiang F, Xu S, Chu X M, et al. MCDTB: A macro-level Chinese discourse TreeBank[C]//Proceedings of the 27th International Conference on Computational Linguistics, 2018: 3493-3504.

[14] 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.

[15] Peng N Y, Dredze M. Named entity recognition for Chinese social media with jointly trained embeddings [C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 548-554,

[16] 杨锦锋，关毅，何彬，等. 中文电子病历命名实体和实体关系语料库构建[J]. 软件学报，2016，27(11)：2725-2746.

[17] 单赫源，张海粟，吴照林. 小粒度策略下基于CRFs的军事命名实体识别方法[J]. 装甲兵工程学院学报，2017，31(1)：84-89.

[18] 冯蕴天，张宏军，郝文宁. 面向军事文本的命名实体识别[J]. 计算机科学，2015，42(07)：15-18，47.

[19] 王学锋，杨若鹏，朱巍. 基于深度学习的军事命名实体识别方法[J]. 装甲兵工程学院学报，2018，32(04)：94-98.

[20] Carletta J. Assessing agreement on classification tasks: The Kappa statistic[J]. Computational Linguistics, 1996, 22(2): 249-254.

[21] 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.

[22] Sang K T, Meulder D F. Introduction to the conll-2003 shared task: Language-independent named entity recognition[C]//Proceedings of the 2003 Conference on Natural Language Learning, 2003: 142-147.

[23] Do gan R I, Leaman R, Lu Z. NCBI disease corpus: A resource for disease name recognition and concept normalization[J]. Journal of Biomedical Informatics, 2014, 47: 1-10.

[24] Yang J, Zhang Y. NCRF++: An open-source neural sequence labeling toolkit[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018.

[25] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

冯鸾鸾(1995—)，硕士研究生，主要研究领域为自然语言处理。
E-mail: llfeng23@stu.suda.edu.cn

李培峰(1971—)，博士，教授，主要研究领域为自然语言处理、机器学习。
E-mail: pfli@suda.edu.cn

李军辉(1983—)，通信作者，博士，副教授，主要研究领域为自然语言处理、机器翻译。
E-mail: jhli@suda.edu.cn